The Open University

M140

Introducing statistics

BOOK 3
Hypothesis testing

The Open University

M140

Introducing statistics

Book 3

Hypothesis testing

This publication forms part of the Open University module M140 *Introducing statistics*. Details of this and other Open University modules can be obtained from Student Recruitment, The Open University, PO Box 197, Milton Keynes MK7 6BJ, United Kingdom (tel. +44 (0)300 303 5303; email general-enquiries@open.ac.uk).

Alternatively, you may visit the Open University website at www.open.ac.uk where you can learn more about the wide range of modules and packs offered at all levels by The Open University.

To purchase a selection of Open University materials visit www.ouw.co.uk, or contact Open University Worldwide, Walton Hall, Milton Keynes MK7 6AA, United Kingdom for a catalogue (tel. +44 (0)1908 274066; fax +44 (0)1908 858787; email ouw-customer-services@open.ac.uk).

The Open University, Walton Hall, Milton Keynes, MK7 6AA.

First published 2013. Second edition 2015.

Edited, designed and typeset by The Open University, using the Open University TEX System.

Printed in the United Kingdom by The Charlesworth Group, Wakefield.

ISBN 978 1 4730 0307 1

2.1

Contents

Unit 6

Truancy

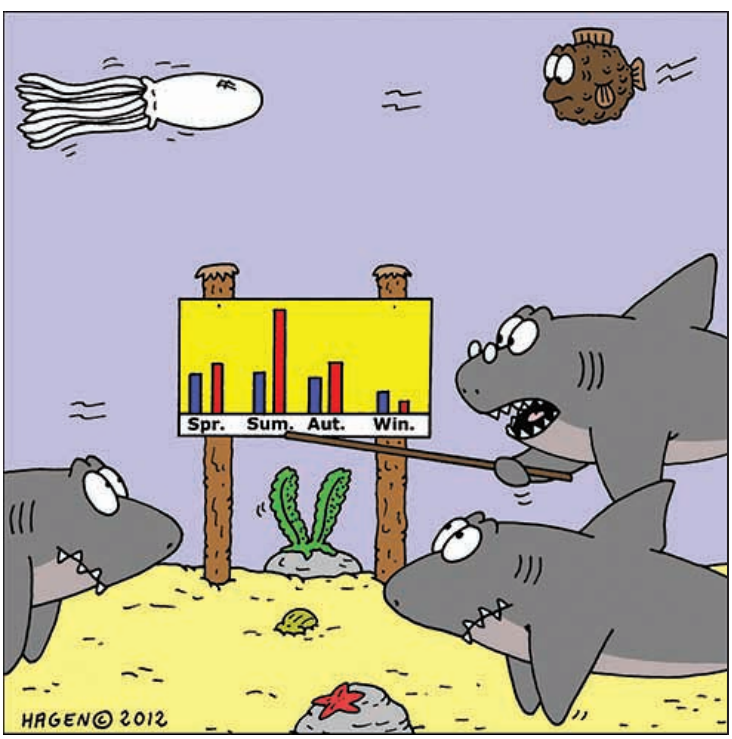
Introduction

In Unit 6 we address the following question:

How often do pupils truant?

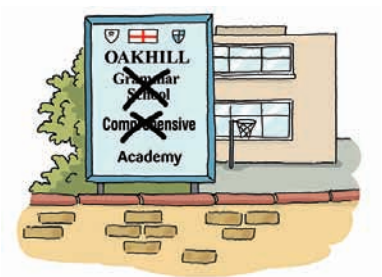
The topic of truancy has been chosen for this unit because it is an example that can be understood without technical background and which enables key statistical ideas to be illustrated and generalised. Statistics can enable informed discussion and decision-making to take place – decisions should be based on the careful analysis of reliable data. This is a module in statistics, not in education, but an aim in this unit and Units 7 to 9 is to demonstrate that statistics has an important role to play in the interpretation of data on education.

The emphasis in this unit is on using statistical techniques to reach conclusions; however, background factors which may account for the conclusions are also discussed. This will give you the opportunity to think about the issues involved, perhaps with particular reference to your own, or your friends', experiences.



Research shows that while the number of surfers is somewhat constant over the year, there is a sudden increase in casual bathers over summer.....

It is worth mentioning that approaches to education in the UK change regularly. This module was written in 2012–13, and if you are studying M140 some years later, the system may have a different structure, and different issues may be important. However, the same statistical techniques will be valid, and many of the questions will still be relevant.



Schools in England

The school system in England appears to be in a continual state of change. In the 1960s and 1970s, most grammar schools and secondary modern schools were replaced by comprehensive schools. Some City Technology Colleges started in the late 1980s and 1990s – they typically specialised in technology, science and mathematics, and often forged close links with businesses and industry.

More recently, many schools have been changing their status to become academies, thus gaining greater independence, particularly in respect of budget control. In 2010 there were only 200 academies, but the number had risen to 1635 by March 2012. Also, free schools were introduced in 2011 as an extension of the Academies Programme. These schools are funded by the taxpayer, but they are not controlled by the local authority and may be set up by parents, teachers, charities and businesses.

Unit 6, along with Units 7, 8 and 9, is a little more mathematical than earlier units. Do not worry if you find them difficult. The important thing is to ensure that you understand the results and how to apply them. In some cases, explanations are given of how results are derived. These are provided as you may find them interesting and they could increase your understanding. However, you will not be expected to reproduce these explanations in detail.

This unit contains six sections. In Section 1, we consider what we mean by the question *How often do pupils truant?*. How are we going to measure truancy? Which pupils are we talking about? We shall narrow the question down until we have a more specific one that we are able to answer. In Sections 2 and 3 we introduce the concept of *probability*, which is fundamental to the ideas of statistical inference. In Sections 4 and 5 we describe a strategy for drawing inferences about a population from a sample, and introduce a particular method, known as the *sign test*, that follows this strategy. Section 6 directs you to the Computer Book, where you will use Minitab to calculate certain probabilities and perform a sign test.

The unit introduces a lot of new concepts and ideas, some of which you may find difficult at first. However, you will find that these concepts occur many times in the remainder of the module, so you will have plenty of opportunity to become familiar with them.

Writing down numbers

In doing calculations, you have been encouraged to retain full accuracy in intermediate steps, perhaps writing down numbers *in full* to demonstrate that you are doing so. This is laborious and makes errors in copying down a number more likely. In the remaining units of this module, you need not do so. As far as you can, you should still use your calculator memory to store numbers to the calculator's accuracy, but you need not write down all numbers in full (usually about five significant figures will be enough, enabling you to focus on the most important part of a number).

1 Clarifying the question

In Section 1 we consider what is meant by the question

How often do pupils truant?

Notice that this question refers implicitly to whole populations: for example, all schools in a particular area. We are usually not directly interested in how the children in one particular school behaved. However, it is often impossible, or at least not feasible, to collect data from the whole population. Instead we select a random sample of data. It might be a random sample of schools or of children. The sample is analysed by the methods we learned in earlier units, and we then need to decide how the results obtained from the sample apply to the whole population.

Random sampling was discussed in Subsection 1.2 of Unit 4.

Statistical inference makes inferences about a population on the basis of data drawn from that population.

The above question about truancy may well have arisen from more general questions, such as:

Why do some pupils learn very little? Are we using good ways of teaching? Does the quality of my child's education depend on where I live?

However, these latter questions can only be tackled if they are first made more precise. Hence, rather than simply *posing* a question, we will often need to *clarify* it, and we may need to clarify it more than once as we learn more about the problem. In earlier units we used the modelling diagram shown in Figure 1 as a framework for how we explore and summarise batches of data.

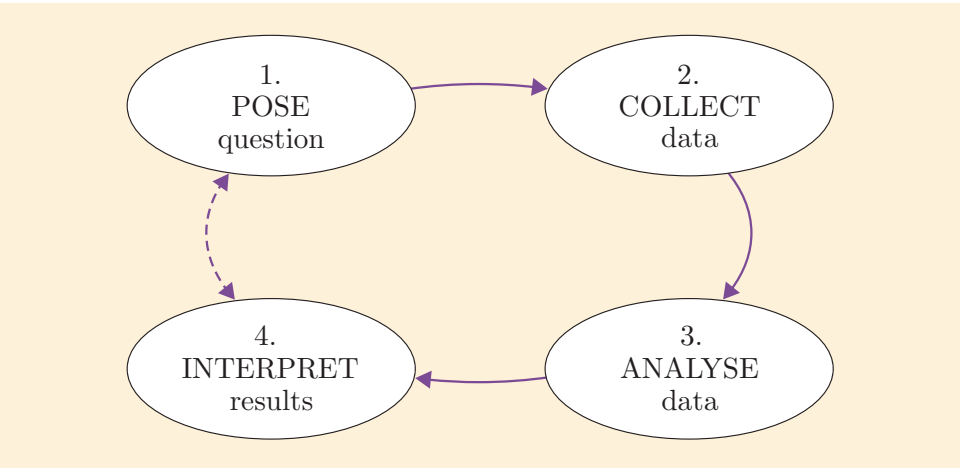


Figure 1 Modelling diagram

From now on we shall use a slightly modified form of this diagram in which

- the differing roles of populations and samples are identified
- the first box is changed from *pose question* to *clarify question*.

The modified modelling diagram is given in Figure 2.

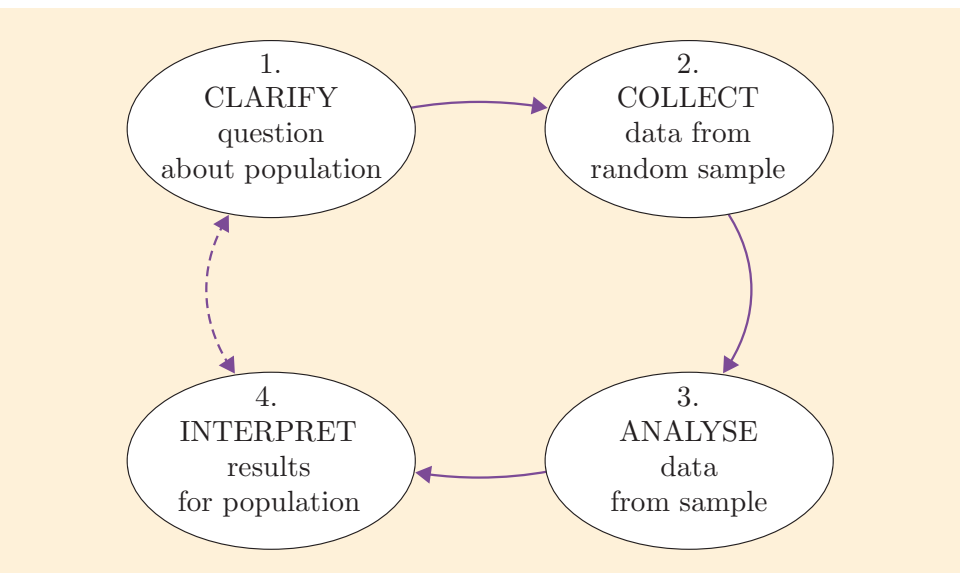


Figure 2 Modified modelling diagram

The processes involved in statistical inference are very important. While we are concentrating on them, we will be concerned particularly with stages 3 and 4 of the modelling diagram. However, this does not mean that we can ignore the other stages. It is always important to ensure that the question under discussion is defined carefully and precisely. Also, although in this unit we shall not be particularly concerned with *how* the data were collected, we do need to know the *form* of the data, as this affects the way the sample is analysed.

1.1 The question to be clarified

As we saw in earlier units, statistics is good at answering questions that require a numerical answer. However, the question for this unit is a very vague question. For example, we might be interested in how much particular children truant, or we might want to compare truancy at different schools.

First, suppose we were looking at particular children.

Activity 1 *Factors affecting a child's truancy*

Spend a few minutes thinking about what factors might affect how much a child plays truant. Then write down three factors that you think might be relevant.

Suppose a child psychologist is helping a particular child with a truancy problem. The psychologist would want to know the child's attendance record and factors about the child's circumstances that can influence truancy. The psychologist would then consider these factors and see if any pattern from the attendance record supported a given factor.

The same approach is followed if you move from considering individual truancy to truancy associated with different schools. We shall concentrate on looking at patterns with regard to schools, not individual children.

There are many different schools, and the amount of truancy will vary greatly. One of the interesting questions is whether different types of school have different amounts of truancy.

Activity 2 *Factors affecting truancy in a school*

Write down three factors that might affect the amount of truancy in a school.

Age of children is one of the most important factors in truancy figures. There is much less truancy at primary schools than at secondary schools. Young children are more likely to be taken to school by their parents, and also, since they are usually with the same class teacher all the time, truancy would be more easily noticed and could be followed up more quickly. We shall concentrate on secondary schools.

As you saw in the solution to Activity 2, there are still many factors that may affect truancy rate even after we have allowed for age to some extent by looking only at secondary schools. They include type of school, location of school and size of school, and there are also other factors, such as the attitude of the teachers, which are more difficult to measure. We shall look at several of these factors in the course of the unit, but we shall start with size of school.

Is the truancy rate in large schools the same as the overall rate?

A definition of 'truancy rate' is given in Subsection 1.2.

To simplify the problem further, and to try to consider a fairly homogeneous group of schools, we shall look at large schools in the East of England region and compare them with all schools in the East of England. It would not be sensible to compare large schools in the East of England with the national average truancy rate, as the East of England has a much lower truancy rate than, for example, Inner London. If we want to investigate the effect of size of school, we want the schools to be alike in *other* respects as much as possible, and location is one way of achieving this. So we shall ask the following question.

Do large secondary schools in the East of England have the same truancy rate as all secondary schools in the East of England?

Changes in pupil performance – up or down?

Pupils in the UK are steadily doing better when compared with their predecessors, while simultaneously doing worse when compared with the rest of the world. On the one hand, in 2011 the overall pass rate of A levels (qualifications typically used for university entrance in England, Wales and Northern Ireland) increased for the 29th consecutive year and more than a quarter of entries achieved the highest grade. Also in 2011, pass rates for Highers (qualifications typically used for university entrance in Scotland) hit a record high. On the other hand, Britain's position in international education ratings dropped in most subjects after 2000.

The OECD (Organisation for Economic Co-operation and Development) publishes reports on economic and social factors in its member states, including school performance league tables. Between 2000 and 2006, the UK fell from 4th in the world to 14th in science, from 7th to 17th in literacy, and from 8th to 24th in mathematics. In part, no doubt, this reflects an increasing importance that other countries are placing on education. As Arne Duncan, the United States Secretary of Education, noted in a speech to UNESCO (United Nations Educational, Scientific and Cultural Organisation), '... in a knowledge economy, education is the new currency by which nations maintain economic competitiveness and global prosperity.'

1.2 Analysing the data

We have now decided on a specific question to investigate and we need to collect some data. This unit is mainly about analysing data that have already been collected, but it is worth spending a little time thinking about exactly what data should be collected. We can assume that we have a sampling frame consisting of all state-funded secondary schools in the East of England and the number of pupils they have. We can therefore pick out large schools, and we have arbitrarily defined these to be schools with 1000 or more pupils. We can then select a random sample of these schools. A sample size of 12 has been chosen.

A sampling frame was defined in Subsection 4.6 of Unit 4 as a list of all individuals in the target population.

We now want a single number to summarise the amount of truancy for each of the 12 schools. First we must consider what we mean by *truancy*. If a child skips school to go to the shops, then they are playing truant, while if they miss school because they are ill in bed, then they are not.

Activity 3 *What is truancy?*

Write down three reasons that a child might miss school – one that is definitely truancy, one that is definitely *not* truancy, and one that might or might not be truancy, depending on circumstances.

A school in Colorado photographed in 1915 during the season to harvest beet – only five pupils are at school while another thirty-five are absent because they are helping with the beet work.

A clear definition of truancy is needed if we are to gather truancy data for the different schools. The definition must take account of what data can be gathered, otherwise the definition may not be useful. In the next activity you are asked to think about how data related to truancy might be collected and used.

Activity 4 *Measures of truancy*

Think of two possible ways in which data on truancy in a particular school could be collected and truancy in the school measured. They should be feasible methods which will not occupy too much of the teachers' time.

When gathering data, precise definitions are needed. Hence the UK government collects data, not on truancy, but on 'unauthorised absence from school'. An unauthorised absence is absence without permission from a teacher or other authorised representative of the school. Records are kept of when permission for absence has been given (which would be retrospectively in the case of illness), so unauthorised absence is a well-defined, documented quantity. It is clearly closely related to truancy. Indeed, when the government publishes statistics on unauthorised absences from school, television and newspapers refer to them as **truancy rates**. We shall do the same.

Truancy rate

One of the statistics on schools that the government publishes is the **unauthorised absence rate**, and we will adopt this as our truancy rate.

- A *pupil's truancy rate* is the proportion of school half-days that the pupil was absent without authorisation.
- A *school's truancy rate* is the average truancy rate of its pupils.

You may well think this is not an ideal measure, and we shall return to this point in Subsection 5.3, but at present we shall concentrate on the analysis.

2. Collect

The module team has used data that were already published. The data give truancy rates over the first two terms of the 2010/2011 school year. The percentage truancy rates in a sample of 12 large schools in the East of England were as follows. (Note that from here on, 'school' means 'secondary school', unless otherwise qualified.)

Table 1 Truancy rates (%) in 12 large schools in the East of England

0.83	1.09	1.84	1.88	1.52	2.78	0.31	1.06	2.90	1.19	1.44	0.82

(Data source: Department for Education (2011) *Pupil absence in schools in England, autumn term 2010 and spring term 2011*)

The median truancy rate for all secondary schools in the East of England during the first two terms of 2010/2011 is known to be 0.98%.

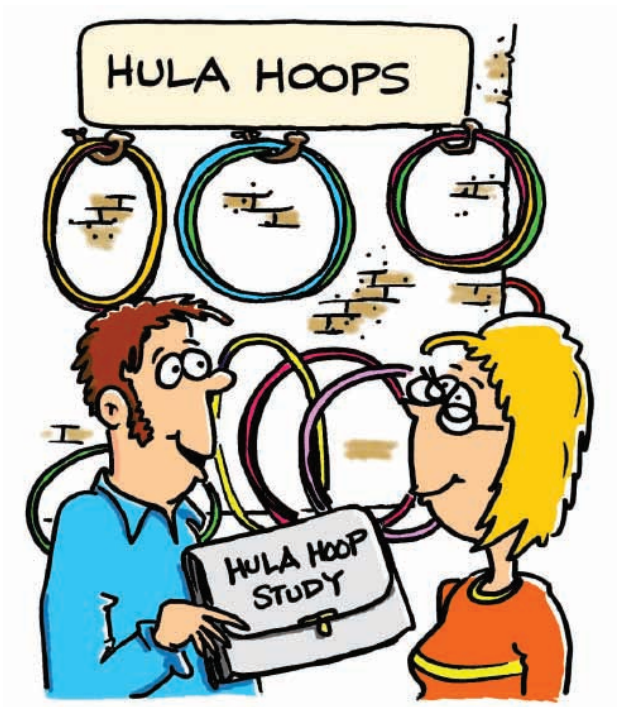
This figure of 0.98% is a population median; see Subsection 3.2 of Unit 4.

We can now move on to stage 3 and then stage 4 of the modelling diagram. We can analyse our sample and then try to interpret our results in terms of the whole population. This last step is known as *statistical inference*.

Statistical inference

This is a process of inferring back from the sample to the population. Somewhat paradoxically, the best approach to making inferences about populations from a sample is to consider what samples are likely to be obtained from a population.

To carry out statistical inference successfully, it is important that we have selected a *random* sample.



'The population of hula hoops tells you what kind of hula hoops will be in your sample. But the sample of hula hoops is used to infer things about the population. Isn't that circular logic?'

Activity 5 *Median truancy rate of sample of 12 large schools*

(a) Find the median truancy rate of the sample of 12 schools.

(b) Can you infer back from the sample to the population of all the East of England schools? In particular, is it possible to infer that the population median of all large schools in the East of England is, or is not, equal to 0.98%?

We do not expect you to give a definitive answer to part (b). Just think a little about the questions asked.

The sample median rate of 1.32 is larger than 0.98 but not by an enormous amount. If we had chosen a different random sample, the median would probably have been different. Here is another random sample of size 12 of truancy rates in large schools in the East of England. This time the values have been written in ascending order:

0.59, 0.67, 0.68, 0.94, 1.35, 1.38, 1.48, 1.54, 1.61, 1.86, 1.89, 1.99.

3. Analyse

In this case the sample median is $\frac{1}{2}(1.38 + 1.48) = 1.43$. This is also bigger than 0.98 and a little larger than the sample median found in Activity 5.

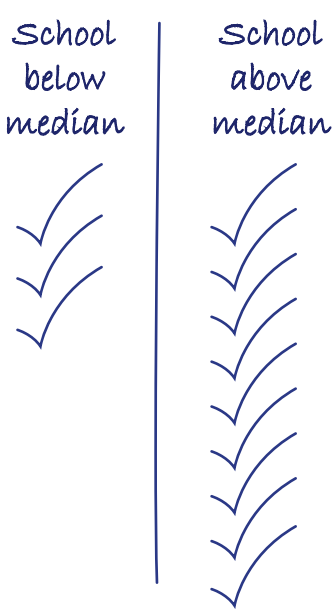
We should not infer back simply by looking at the sample and guessing. We need a systematic method, and the development of such a method is the main purpose of this unit.

4. Interpret

We shall look at the data of Table 1 in a different way. Instead of considering the sample median, we shall take each value in turn and record whether it is larger or smaller than 0.98. The first value of 0.83 is below 0.98, so we shall record that as a negative difference, or [−]. The value 1.09 lies above 0.98, so we record it as [+]. Working through the whole sample, we obtain:

0.83	1.09	1.84	1.88	1.52	2.78	0.31	1.06	2.90	1.19	1.44	0.82
[−]	[+]	[+]	[+]	[+]	[+]	[−]	[+]	[+]	[+]	[+]	[−]

Altogether, we have nine values above 0.98 and three values below it.



In Section 4, we are going to assume that the median percentage truancy rate for large schools in the East of England is indeed 0.98. We are then going to try to answer the question:

> *How likely is it that a random sample of size 12 would contain nine values above and three values below the assumed population median?*

To answer this question, we need a way of making an informed judgement.

If the median truancy rate for large schools were 0.98, it would not be surprising if there were six values above 0.98 and six values below it. This is the result that seems most likely, since the definition of a population median is that half the population values lie above it and half lie below it. We should expect a representative sample to be similar to the population. Also, after our work with samples in Unit 4, we should not be very surprised to find seven values above the median and five below it (or the other way round).

However, it would seem surprising if all 12 values were either above or below the population median. So if all 12 were greater than 0.98 (the assumed population median), we would suspect that the population median for large schools was larger than 0.98.

This suggests that some events are more likely to occur than others. The next activity suggests you think further along these lines.

Activity 6 *Which events are more likely to happen?*

Look at the following list of events and rank them in order of likeliness to occur, starting with the one you think most likely.

A. Your new colleague at work has the same birthday as you.

B. Two out of a group of ten people have the same birthday.

C. The sun will rise tomorrow.

D. The sun will shine tomorrow.

E. You will win the jackpot in the National Lottery next week.

F. You toss a coin and it lands tails up.

G. You throw a die and it shows a six. (Remember 'die' is the singular form of 'dice'.)

H. A member of a hockey team is 150 years old.

Activity 6 has demonstrated that some events are certain to occur, others are impossible, and the likelihoods of occurrence of others are somewhere between these two extremes. In the next two sections, we shall introduce the concept of probability, which is a method of associating a numerical measure with the likelihood of occurrence of a particular event.

We shall then be able to use this concept to calculate how likely we are to observe nine values above and three values below the population median in a random sample of size 12. Then we shall be in a better position to use the sample to make inferences about the population and decide whether or not the median truancy rate of the population of large secondary schools in the East of England during the first two terms of 2010/2011 is equal to 0.98%.

Exercises on Section 1

Exercise 1 *Your view on the likely occurrence of more events*

Look at the following list of events and rank them in order of likeliness to occur, starting with the one you think most likely.

A. A husband and wife find they were born on the same day of the week.

B. England will win the next football World Cup.

C. A Brazilian team will win the next football European Cup.

D. Death and taxes.

E. A mother's first pregnancy results in twins.

F. You get exactly three heads when you toss a coin five times.

G. A chicken egg has two yolks.

H. It snows in London on Christmas day.

I. You will be struck by lightning next year.

2 Probability

At the end of Section 1, we established it would be useful to associate a number with an event so that it provided a measure of the likelihood of that event. This number is called the **probability** of occurrence of the event. The concept of probability plays such an important role in statistics that we are going to spend two sections investigating its properties, before returning to our specific question of truancy.

'How do you want it – the crystal mumbo-jumbo or statistical probability?'

2.1 Measuring chance

The easiest way of thinking about probability is to equate it to proportion: the *probability* that a particular event will happen is the *proportion* of the time that it is expected to happen. When we toss a fair (unbiased) coin, for example, there is a fifty-fifty chance that the coin will land 'heads' because half the time it should land 'heads' and half the time it should land 'tails'. That is, the proportion of time that the outcome should be heads is $\frac{1}{2}$, so the *probability* that the outcome will be heads is $\frac{1}{2}$.

(In practice, of course, you can only toss a coin a finite number of times, and it is very unlikely to land 'heads' exactly half the time. For example, if you toss it 600 times, then there is little chance that it will land 'heads' exactly 300 times. However, if you toss a coin an enormous number of times, the proportion of 'heads' should be very close to $\frac{1}{2}$.)

Similarly, if a die is fair, then each of its six sides is equally likely to be the outcome when it is rolled. Thus, for example, the proportion of rolls that should result in a 4 is $\frac{1}{6}$, so the probability of rolling a 4 is $\frac{1}{6}$.

Probability = Proportion

See Subsection 1.2 of Unit 4.

You met the ideas of random selection and random sampling in Unit 4. With random sampling, each member of the population is equally likely to be included in the sample. In particular, if a person or item is picked at random from a population, each member of the population is equally likely to be the one that is picked. We shall use these ideas to begin our investigation of probability.

Example 1 *Student probabilities*

Table 2 gives the breakdown of male and female students in a small university, by their faculty. (A student belongs to exactly one faculty.)

Table 2 Student population of a small university

	Science	Arts	Law	Medicine	Total
Female	964	1532	87	206	2789
Male	1638	1039	247	369	3293
Total	2602	2571	334	575	6082

Now suppose that a student is selected at random from the university. Using the notion that probability equates to proportion, various probabilities relating to the student can be calculated. For example, the table gives the number of students who are female (2789) and the total number of students (6082). Hence we can determine the probability that the selected student is female by calculating the proportion of students who are female, as follows.

The probability of selecting a female student is

$$\frac{\text{number of female students}}{\text{total number of students}} = \frac{2789}{6082} \simeq 0.459.$$

Similarly, the probability of selecting an arts student is

$$\frac{\text{number of arts students}}{\text{total number of students}} = \frac{2571}{6082} \simeq 0.423,$$

and the probability of selecting a male science student is

$$\frac{\text{number of male science students}}{\text{total number of students}} = \frac{1638}{6082} \simeq 0.269.$$

You have now covered the material related to Screencast 1 for Unit 6 (see the M140 website).

Activity 7 *Student probabilities*

Giving your answers to three decimal places, what is the probability that a student chosen at random from the university population in Table 2 is:

(a) a male student?

(b) a law student?

(c) a female medical student?

Whenever you perform a calculation, you should look to see if your answer seems sensible – that way, you can sometimes spot large calculation errors. In the last activity you would expect your probability for (a) to be large because a lot of the students are male, while the probabilities for (b) and (c) should be small because there are comparatively few law students or female medical students.

Activity 8 *More probabilities*

Suppose we have a population of men and women and select just one person at random. For each of the following cases, give the probability that a woman is selected.

(a) The population of size 10 consists of 5 men and 5 women.

(b) The population of size 10 consists of 1 man and 9 women.

(c) The population of size 10 consists of 9 men and 1 woman.

(d) The population of size 10 consists of 10 men and 0 women.

(e) The population of size 100 consists of 99 men and 1 woman.

(f) The population of size 100 consists of 0 men and 100 women.

In part (d) of Activity 8, the probability of selecting a woman was 0. This particular population contained no women, so it was impossible to select a woman. Similarly in part (f), the probability of selecting a woman was 1, and in that situation, the population contained no men, so a woman was certain to be selected.

If you look at the other probabilities we have calculated, you will see that they are all between 0 and 1. This is because they have all been calculated as the proportion of a population who possess a certain property – for example, being a science student or being a woman. The closer the probability is to 1, the more likely the event is; the closer the probability is to 0, the rarer the event is. If you ever calculate the probability of an event and find it is either greater than 1 or a negative number, then you know you have made a mistake somewhere.

This discussion can be summarised by the following three important properties of probability:

- If an event is impossible, then its probability is 0.
- If an event is certain, then its probability is 1.
- Any event which is uncertain but not impossible has a probability that lies between 0 and 1.

At this point it is convenient to introduce some notation which cuts down the number of words we have to write. In the last activity you looked at the probability of the event that a person selected at random from a population is a woman. If we let W stand for the event *a woman is selected*, then we shall write $P(W)$ for the probability that event W occurs, or the probability that a woman is selected. So

Both P and Pr are commonly used to mean 'probability' – we shall use P.

$$P(W) = \frac{\text{number of women in population}}{\text{total number in population}}.$$

Similarly, we might let L stand for the event that a student selected at random is a law student. Then

$$P(L) = \frac{\text{number of law students in population}}{\text{total number in population}}.$$

This can be extended to a more general situation.

Let E stand for the event of selecting a person or object with some particular property from a population, using random sampling. Then the **probability** of E is given by

$$P(E) = \frac{\text{number in population with particular property}}{\text{total number in population}}.$$

2.2 Adding probabilities

Probability gives a way of measuring how likely an event is to occur in random sampling. In the last subsection you learned that a probability is always greater than or equal to 0 and always less than or equal to 1. The following example and activity use data on truancy to help you become more familiar with the idea of probability. The same data is then used to explore another property of probability.

Example 2 *Truancy in two schools*

Table 3 shows some invented data on truancy in two schools, A and B, that contained 200 and 100 pupils, respectively.

We have invented the data so that the numbers are small and the calculations are easy to check.

Table 3 Truancy numbers in two schools

Number of days absent	Number of pupils School A	School B	Total numbers
0 to 4	108	42	150
5 to 9	60	30	90
10 to 19	26	19	45
20 or more	6	9	15
Total	200	100	300

We shall use the table to answer the following questions.

1. If a child is selected at random from these two schools, what is the probability that this child was absent through truancy for fewer than 5 days?
2. If a child is selected at random from these two schools, what is the probability that this child is at School A and was absent through truancy for fewer than 5 days?
3. If a child is selected at random *from School A*, what is the probability that this child was absent through truancy for fewer than 5 days?

Let T stand for the event that a child selected at random was absent through truancy for fewer than 5 days, and let A stand for the event that the child attends School A.

1. Here the probability is $P(T)$. Now

$$P(T) = \frac{\text{total number of children absent for} < 5 \text{ days}}{\text{total number of children}} = \frac{150}{300} = 0.5.$$

The symbol $<$ means 'less than'. Similarly, $>$ means 'greater than', $\leq$ means 'less than or equal to' and $\geq$ means 'greater than or equal to'.

So there is a probability of 0.5 that a child picked at random from these two schools was absent through truancy for fewer than 5 days.

2. Here the probability is that both events T and A occur. This is $P(T \text{ and } A)$, which is an extension of our notation for the probability of an event. (It means the probability that both T and A occur. In this case, the event 'T and A' occurs if a child is absent through truancy for fewer than 5 days and also attends School A.) From Table 3, we see that 108 children attended School A and were absent through truancy for fewer than 5 days. So

$$P(T \text{ and } A) = \frac{\text{total number of children satisfying both } T \text{ and } A}{\text{total number of children}} = \frac{108}{300} = 0.36.$$

So the probability that a child attends School A and is absent through truancy for fewer than 5 days is 0.36.

3. You have to be careful here to appreciate what probability is required and how it differs from part (b). Probability is giving us precise answers, so we have to be careful that we ask precise questions and that the questions are the ones we want to answer. Here we assume that the child is selected at random from those at School A, so School A provides the total population. Thus

$$P(\text{child from School A is absent for} < 5 \text{ days}) = \frac{\text{total number of children at School A absent for} < 5 \text{ days}}{\text{total number of children at School A}} = \frac{108}{200} = 0.54.$$

So the required probability is 0.54.

Activity 9 *Truancy in two schools*

(a) Find the probability of each of the following events, using the data in Table 3.

- A child chosen at random from these two schools is absent through truancy for between 5 and 9 days.
- A child chosen at random from these two schools is absent through truancy for between 5 and 9 days and attends School B.
- A child chosen at random from those attending School A is absent through truancy for between 10 and 19 days.
- A child chosen at random from these two schools is absent through truancy for 10 or more days.

(b) For the first two events in part (a), define letters to denote the appropriate events as in Example 2. Use these letters to write in $P(E)$ notation the probabilities you calculated.

We next consider when probabilities can be added together. In Activity 9(a) you calculated the probability that a child selected at random from the two schools is absent for 10 or more days. This probability is 0.20. Now absence for 10 or more days means either absence for 10 to 19 days or absence for 20 or more days; in the context of the question, it does not matter which of these two events actually occurs.

We can easily find the probability of occurrence of each of these events:

$$P(\text{child absent for 10 to 19 days}) = \frac{45}{300} = 0.15,$$

$$P(\text{child absent for 20 or more days}) = \frac{15}{300} = 0.05.$$

The sum of these two probabilities is $0.15 + 0.05 = 0.20$, which is the probability that a child selected at random is absent for 10 or more days.

So we have

$$P(\text{child absent for} \geq 10 \text{ days}) = P(\text{child absent for 10 to 19 days}) + P(\text{child absent for} \geq 20 \text{ days}).$$

This is clearly not a coincidence, since

$$\begin{aligned} &P(\text{child absent for} \geq 10 \text{ days}) \\ &= \frac{\text{total number of children absent for} \geq 10 \text{ days}}{\text{total number of children}} \\ &= \frac{\text{number absent for 10 to 19 days} + \text{number absent for} \geq 20 \text{ days}}{\text{total number of children}} \\ &= \frac{\text{number absent for 10 to 19 days}}{\text{total number of children}} + \frac{\text{number absent for} \geq 20 \text{ days}}{\text{total number of children}} \\ &= P(\text{child absent for 10 to 19 days}) + P(\text{child absent for} \geq 20 \text{ days}). \end{aligned}$$

This example was about two events, either of which might occur. However, it is impossible that both events occur at the same time. A child cannot both be absent for 10 to 19 days and for 20 or more days. Any two events with this property are called *mutually exclusive*. (The two events 'absent for 10 or more days' and 'absent for 20 or more days' are not mutually exclusive. If a child were absent for 21 days, both of these events would occur.)

Two events are said to be **mutually exclusive** if they cannot occur at the same time. More generally, any number of events are said to be mutually exclusive if no two of them can occur at the same time.

Activity 10 *Exclusive events?*

Which of the following pairs or sets of events are mutually exclusive?

(a) A person's blood type is measured and the events are: (i) it is blood type A, and (ii) it is blood type O.

(b) A person is described and the events are: (i) the person has black hair, and (ii) the person has blue eyes.

(c) Some dice are rolled and the events are: (i) 3 dice are rolled, (ii) each die gives the same number, and (iii) the sum of the numbers on the dice is 8.

(d) A person's blood type is measured and the events are: (i) it is blood type A, (ii) it is blood type O, and (iii) the person's blood is rhesus positive.

Suppose A and B are two mutually exclusive events. Since they are mutually exclusive events, both of them cannot occur at the same time. We shall denote the probability that one of A and B occurs by $P(A \text{ or } B)$. This probability is given by the addition rule for mutually exclusive events.

Addition rule for mutually exclusive events (the 'or' linkage)

For any two mutually exclusive events A and B,

$$P(A \text{ or } B) = P(A) + P(B).$$

This rule extends to more than two events when they are all mutually exclusive. For example, if A, B and C are mutually exclusive events, then

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C).$$

Activity 11 *Adding two probabilities*

This activity relates to the data in Table 3 in Example 2, reproduced below for convenience.

Table 4 Truancy numbers in two schools

Number of days absent	Number of pupils School A	School B	Total numbers
0 to 4	108	42	150
5 to 9	60	30	90
10 to 19	26	19	45
20 or more	6	9	15
Total	200	100	300

(a) Suppose a child is selected at random from School A. Find the probabilities that this child is absent through truancy for

- 0 to 4 days
- 5 to 9 days
- 0 to 9 days.

(b) Verify that the addition rule for mutually exclusive events holds in this case.

There is a particular application of the addition rule which gives us a useful result about probabilities. Suppose E is some event that may or may not occur at a certain trial. (In statistics, a sequence of observations or a sequence of tests is often called a set of *trials*.) For example, the trial might be the selection of one child from the population of the two schools, and E might be the event that the child is absent through truancy for 0 to 4 days. Let F be the event that E does *not* occur. Then F is the event

that a child is absent for 5 or more days. The events E and F are called **complementary**, because only one of them can happen and together they *complete* the list of possibilities.

As one of the events E or F must occur, $P(E \text{ or } F) = 1$. (Think about this. Is it possible that neither E nor F occurs?)

Hence, by the addition rule,

$$P(E \text{ or } F) = P(E) + P(F) = 1.$$

So $P(F) = 1 - P(E)$. We express this rule in words, as follows.

Probability rule for complementary events (the 'not' linkage)

For any event E,

$$P(E \text{ does not occur}) = 1 - P(E \text{ does occur}).$$

You have now covered the material related to Screencast 2 for Unit 6 (see the M140 website).

Activity 12 *A small university revisited*

Table 2 (reproduced below as Table 5) gave the breakdown of male and female students in a small university, by their faculty.

Table 5 Student population of a small university

	Science	Arts	Law	Medicine	Total
Female	964	1532	87	206	2789
Male	1638	1039	247	369	3293
Total	2602	2571	334	575	6082

Suppose that a student is selected at random.

(a) Find the probability that the student is male. Hence determine the probability that the student is female.

(b) What is the probability that the student is not a medical student?

Activity 13 *Occasional truant, or not*

In relation to the data in Table 3 (reproduced in Table 4), let G be the event that a child selected at random from the two schools is absent through truancy for 5 to 9 days.

(a) Describe the event 'G does not occur'.

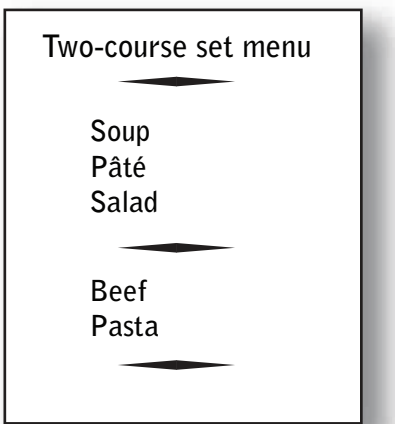
(b) Use the above rule to evaluate: $P(G \text{ does not occur})$.

2.3 Multiplying probabilities

We have seen that probabilities are added when we have the 'or' linkage, and want P(A *or* B). We next consider how to determine probabilities when we have an 'and' linkage, and want P(A *and* B). We use the notion that P(A and B) is the proportion of the time that A and B both happen.

Example 3 *Two-course lunch*

A restaurant offers a two-course set lunch. There are three choices for the first course – soup, pâté or salad – and two choices for the second course – beef or pasta. The different meal-combinations are shown in Figure 3.



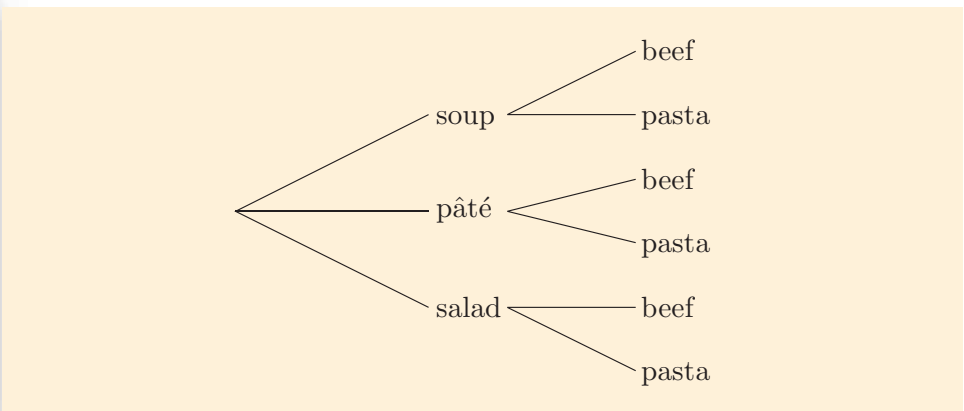


Figure 3 Tree diagram for a two-course lunch

The diagram in Figure 3 is referred to as a **tree**. Starting at the left of the figure, we can follow one of three lines – **branches** – to choose a first course (soup, pâté or salad). From each first course we can follow one of two lines – **sub-branches** – to choose the second course (beef or pasta). Thus there are $3 \times 2 = 6$ different paths we can follow, corresponding to the six possible meal combinations: soup–beef, soup–pasta, pâté–beef, pâté–pasta, salad–beef and salad–pasta.

Suppose, now, that we choose a first course at random and also choose the second course at random. Then each of these six possibilities is equally likely. Thus the proportion of time we choose, say, salad followed by beef would be one-sixth, so

$$P(\text{salad and beef combination}) = \frac{1}{6}.$$

Notice that there is a choice of three first courses, so if the choice is made at random,

$$P(\text{salad for first course}) = \frac{1}{3}.$$

And, as there are two choices for the second course,

$$P(\text{beef for second course}) = \frac{1}{2}.$$

Consequently, in this example

$$P(\text{salad and beef combination}) = P(\text{salad}) \times P(\text{beef}).$$

Extending Example 3 is helpful, so suppose that there are four choices for the first course – soup, salad, pâté and prawns – and five choices for the second course – beef, chicken, fish, pasta and quiche. Following similar reasoning to Example 3, there are $4 \times 5 = 20$ different meal combinations.

Activity 14 *Two-course meal*

(a) Draw a diagram similar to Figure 3 to show the 20 different meal combinations.

(b) Suppose the soup and salad are both vegetarian options for the first course, and pasta and quiche are vegetarian options for the second course. Write down the different combinations of courses that a vegetarian could have.

(c) If a first course and second course are selected at random, what is the probability that the combination is vegetarian?

(d) Suppose a first course and second course are selected at random.

- What is the probability that the first course is vegetarian?
- What is the probability that the second course is vegetarian?

Check that the multiplication of these answers gives the probability you found in (c).

In Activity 14 you found that

$$P(\text{vegetarian first course } \textit{and} \text{ vegetarian second course}) = P(\text{vegetarian first course}) \times P(\text{vegetarian second course}).$$

The reason is as follows:

$$\begin{aligned} &P(\text{vegetarian first course } \textit{and} \text{ vegetarian second course}) \\ &= \frac{\text{number of vegetarian meal combinations}}{\text{total number of meal combinations}} \\ &= \frac{\text{number of vegetarian first courses} \times \text{number of vegetarian second courses}}{\text{total number of first courses} \times \text{total number of second courses}} \\ &= \frac{\text{number of vegetarian first courses}}{\text{total number of first courses}} \times \frac{\text{number of vegetarian second courses}}{\text{total number of second courses}} \\ &= P(\text{vegetarian first course}) \times P(\text{vegetarian second course}). \end{aligned}$$

To extend this result, the following activity concerns three-course meals.

What's the probability of this meal?

Activity 15 *Three-course meal*

Suppose the menu for a three-course meal has the same choices of first course and second course as in Activity 14 and it additionally has a third course with three options: pie, crumble or ice cream.

(a) How many different three-course meal combinations are there?

(b) Suppose the three courses are each selected at random.

- What is the probability that the third course would be good with custard (i.e. pie or crumble)?
- What is the probability that the first two courses are vegetarian and the third course is good with custard?

(c) When the three courses are selected at random, check that

$$\begin{aligned} &P(\text{vegetarian first course } \textit{and} \text{ vegetarian second course} \\ &\quad \textit{and} \text{ third course is good with custard}) \\ &= P(\text{vegetarian first course}) \times P(\text{vegetarian second course}) \\ &\quad \times P(\text{third course is good with custard}). \end{aligned}$$

We now formulate in general terms the rule we have been using which enables us to multiply probabilities in appropriate circumstances. To do this we need the concept of 'statistical independence'.

Two events are said to be **statistically independent** if the occurrence of one has no effect on the likelihood of occurrence of the other.

In the last activity, for example, we assumed that each course was chosen at random, so the choice made for one course had no influence on the choice made for any other course. Thus, for example, the two events 'choosing beef for the second course' and 'choosing crumble for the third course' are statistically independent.

Activity 16 *Independent events?*

Which of the following pairs or sets of events are independent?

(a) Picking a person at random where the events are: (i) the person is taller than average; (ii) the person is heavier than average.

(b) Picking two people at random where the events are: (i) the first person is taller than average; (ii) the second person is heavier than average.

(c) Picking a person at random where the events are: (i) the person is taller than average; (ii) the person was born on a Tuesday.

(d) Tossing two coins where the events are: (i) the first coin lands 'heads'; (ii) the second coin lands 'heads'; (iii) both coins give the same outcome.

(e) Drawing a card from an ordinary deck of playing cards where the events are: (i) the card is an ace; (ii) the card is a diamond.

Let A and B be two statistically independent events, and let 'A and B' denote the event that both A and B occur. Then the probability that A and B both occur is given by the following rule.

Multiplication rule for statistically independent events (the 'and' linkage)

If A and B are statistically independent events, then the probability that A and B both occur is given by

$$P(A \text{ and } B) = P(A) \times P(B).$$

This rule extends to more than two events. For example, if A, B and C are statistically independent events, then the probability that they all occur is given by

$$P(A \text{ and } B \text{ and } C) = P(A) \times P(B) \times P(C).$$

You have now covered the material related to Screencast 3 for Unit 6 (see the M140 website).

Earlier in this section, we determined probabilities when an individual was picked at random from a population. In practice, often we randomly sample a number of individuals or a number of items from a population. We shall use the *addition rule* and *multiplication rule* to look at probabilities for samples that do not consist of just one individual but of several. First we will look at probabilities for samples of size 2, before seeing how our results can be generalised to larger samples.

When we introduced random samples in Unit 4, we saw how they could be selected using random number tables. We selected a random number to give us the first member of the sample, then a second, and so on. If the same random number appeared a second time, we rejected it, because the same individual cannot appear more than once in a random sample. For the moment, we shall relax this restriction, because it makes the probabilities easier to find. Thus, if an individual or item is selected for a random sample, it can be picked again at a second selection. Later in this section, we shall return to proper random sampling and show that in practice for large populations, it makes hardly any difference to the probabilities.

We shall introduce the method of finding probabilities for a sample of size 2 by means of an example.

Example 4 *Sample of two*

Suppose we have a population of size 10 which contains three women and seven men. We shall call the three women Ashia, Brenda, and Clare, and the seven men David, Ejike, Frank, Gavin, Harry, Ian and Joe, and refer to them all by their initials. Suppose we select a sample of size 2 from this population, remembering that we may select the same person twice.

How many samples of size 2 are there altogether? We must differentiate between the person selected the first time and the person selected the second time. We might pick Joe first and Clare second; we could write this as (J, C). Alternatively, we might pick Clare first and Joe second; this is (C, J). Other samples include (E, F) and (A, A), as Ashia might be selected twice. There are 10 possibilities for the first member of the sample and 10 possibilities for the second member. Altogether there are $10 \times 10 = 100$ possible samples of size 2.

Activity 17 *Sample of two women*

(a) For the population introduced in Example 4, write down all the samples of size 2 that contain two women, using their initials A, B, C. How many are there?

(b) How could you have found this number without writing down all the samples?

So using the results from Example 4 and Activity 17, we can now write down the probability that our sample of size 2 from this population of ten people contains two women.

Equating probability to proportion, we can write

$$P(\text{sample of size 2 contains two women}) = \frac{\text{number of samples of size 2 containing two women}}{\text{total number of samples of size 2}} = \frac{9}{100} = 0.09.$$

We might also obtain this probability by applying the multiplication rule. Now the probability that the first selection is a woman is $3/10 = 0.3$, as is the probability that the second selection is a woman. So to obtain the probability that both selections are women, we multiply the probabilities together. We shall introduce some notation to describe this result. Let W denote the event that at any selection a woman is chosen, and let $2W$ denote the event that two women are selected in a sample of size 2. Then

$$P(2W) = P(W) \times P(W) = 0.3 \times 0.3 = 0.09.$$

We shall continue to sample from the population of Ashia, Brenda, Clare, . . . , Joe.

Activity 18 *Sample of two men*

Suppose M denotes the event that at any selection a man is chosen, and $2M$ denotes the event that two men are selected in a sample of size 2. Use the multiplication rule to find $P(2M)$.

We can use both this multiplication rule and the addition rule from Subsection 2.2 to find the probability that a sample of size 2 contains one woman and one man. To obtain such a sample, we either select a woman first and a man second, or a man first and a woman second. These possibilities are mutually exclusive events, so we add the probabilities:

$$\begin{aligned} &P(\text{sample of size 2 contains one woman and one man}) \\ &\quad = P(\text{a woman is selected first and a man is selected second}) \\ &\qquad + P(\text{a man is selected first and a woman is selected second}). \end{aligned}$$

Now we know that at any selection, $P(W) = 0.3$ and $P(M) = 0.7$. So, just as with a sample of two women, we multiply the probabilities together to obtain each of the probabilities on the right-hand side. Writing $P(1W)$ for the probability that a sample of size 2 contains one woman (and therefore one man), we obtain

$$\begin{aligned} P(1W) &= P(W) \times P(M) + P(M) \times P(W) \\ &= 0.3 \times 0.7 + 0.7 \times 0.3 \\ &= 0.21 + 0.21 = 0.42. \end{aligned}$$

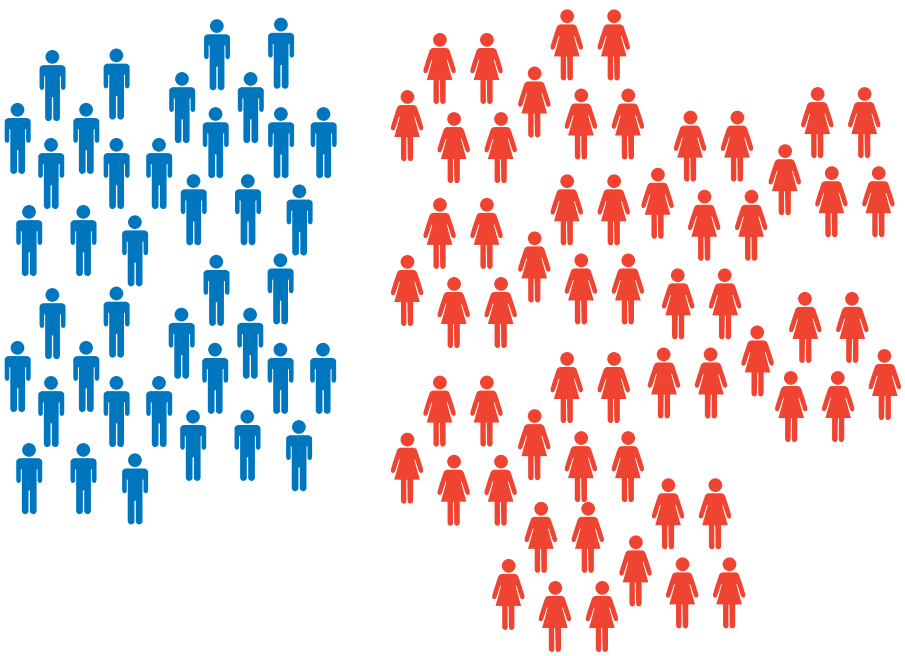
We can now check our results by using the fact that our sample is certain to contain either two women or two men, or one woman and one man; that is,

$$P(2W) + P(2M) + P(1W) = 1.$$

Substituting the calculated values, we obtain $0.09 + 0.49 + 0.42 = 1$, as required.

Activity 19 *Sampling from 100 people*

A population contains 40 men and 60 women. A sample of size 2 is drawn in which the first person selected is also available for selection the second time. Find the probabilities for the different possible numbers of men in the sample, and check that the probabilities sum to 1.



Earlier, we decided to introduce the calculation of probabilities for samples of size 2 by allowing the same individual to be selected twice. Now we shall consider probabilities for random samples of size 2, when any individual can appear only once in a sample. Let us start by looking again at the population of Example 4 (Ashia, Brenda, Clare, ..., Joe), which consists of three women and seven men. We can select the first member of the sample in 10 ways. Then only nine individuals remain, so we can select the second in 9 ways. Hence the total number of samples of size 2 is $10 \times 9 = 90$.

Activity 20 *Sampling two different women*

How many random samples of size 2 contain two different women? Remember that Ashia first and Brenda second is a different sample from Brenda first and Ashia second, and that now the same person cannot be selected twice.

Using the result of Activity 20, we can find

$$P(\text{random sample of size 2 contains two women}) = P(2W) = \frac{6}{90} = \frac{1}{15} = 0.067 \quad \text{(rounded to three decimal places).}$$

Similarly, the probability of there being two men in the sample is

$$\frac{7 \times 6}{90} = \frac{42}{90} = \frac{7}{15} \simeq 0.467.$$

So, by the addition rule and the probability rule for complementary events, the probability of one man and one woman is

$$1 - \left(\frac{1}{15} + \frac{7}{15}\right) = \frac{7}{15} \simeq 0.467.$$

These probabilities – 0.067, 0.467, 0.467 – are different from those we found before when the same individual could be selected twice: 0.09, 0.49, 0.42.

However, this is not a realistic situation. Real-life surveys do not deal with populations of size 10. Suppose our population contains 1000 people (still a fairly small population) of which 300 are women and 700 are men.

Now the probability of selecting a woman is 300/1000, which is still 0.3. So if we *can* select the same person twice, the probability that a sample contains two women is $0.3 \times 0.3 = 0.09$, by the multiplication rule as before. For a random sample in which we *cannot* select the same person twice,

$$P(\text{random sample of size 2 contains two women}) = \frac{300 \times 299}{1000 \times 999} = 0.090 \quad \text{(rounded to three decimal places)}.$$

This is essentially the same as for the simple case when we just multiply 0.3 by 0.3. The probability of two men is $(700 \times 699)/(1000 \times 999) = 0.490$ rounded to three decimal places, again the same as the simple case. This is because with a large population, the probability of selecting the same individual twice is so small that it can be ignored.

For the rest of this module we shall assume that populations are large enough for us to apply the multiplication rule, and that the probability of selecting an individual possessing some characteristic remains the same for each selection of a random sample.

In this section, we have introduced the concept of probability and seen how to calculate probabilities for various events, including the composition of random samples of size 2. In the next section, we shall look at probabilities for larger random samples, and relate these to the question on truancy that we posed in Section 1.

Exercises on Section 2

Exercise 2 *Colouring probabilities*

The hair colour and eye colour of a random sample of 6800 German men are recorded in Table 6.

Table 6 Hair colour and eye colour of 6800 men

		Hair colour				
		Brown	Black	Fair	Red	Total
Eye colour	Brown	438	288	115	16	857
	Grey or green	1387	746	946	53	3132
	Blue	807	189	1768	47	2811
	Total	2632	1223	2829	116	6800

(Source: Goodman, L.A. and Kruskal, W.H. (1954) 'Measures of association for cross classifications', *Journal of the American Statistical Association*, vol. 49, no. 268, pp. 732–64)

Suppose that one of the men in the sample is picked at random. What is the probability that the man:

(a) has blue eyes?

(b) has brown hair?

(c) has blue eyes and brown hair?

(d) has brown hair or black hair?

(e) does not have black hair?

Exercise 3 *Saturday afternoon*

The probability that Sue will go to watch her favourite hockey team play hockey on Saturday is 0.3. The probability that her team will win is 0.4. What is the probability that Sue will watch her team play on Saturday and they will win?

3 Probability distributions from random samples

At the end of Section 1 we began to investigate the median truancy rate of large schools in the East of England by taking a random sample of the schools. In the sample we found that there were nine values above the (assumed) population median, and three values below. This led us to ask the following question.

> *How likely is it that when selecting a random sample of size 12 from some population, the resulting data will have nine values above the (actual) population median and three below it?*

This is really a question about probabilities, and we can answer it by applying the rules for adding and multiplying probabilities that we studied in Section 2. Although the question refers specifically to nine values above and three values below the population median, our approach will enable us to calculate probabilities for all possible outcomes for any size of random sample. To this end, we need other results for counting the number of ways in which an event can occur.

3.1 Counting combinations

Example 5 *Chairperson and secretary*

Suppose a club has ten members and we want to know the number of ways in which a chairperson and a secretary for the club could be chosen, if they cannot be the same person. If the chairperson is chosen first, then there are 10 choices of chairperson. There remain nine people from whom to choose the secretary, so the total number of ways of choosing a chairperson and secretary is

$$10 \times 9 = 90.$$

If the club members were the ten people in Example 4 (Ashia, Brenda, Clare, ...), then the 90 choices would be (A, B), (B, A), (A, C) etc.

Note that it makes no difference whether the chairperson or secretary is picked first. If the secretary were picked first, then there would be 10 choices for the secretary, and nine people would remain from whom to pick the chairperson. Again the number of choices for secretary and chairperson would be $10 \times 9 = 90$.

The Programme Committee for the 2012 European Conference on Quality in Official Statistics going to lunch ... or trying to avoid being appointed as chairperson?

Example 6 *Chairperson, secretary and treasurer*

Suppose the club with ten members want a treasurer as well as a chairperson and secretary. If the chairperson is chosen first, then there are 10 choices of chairperson. If the secretary is chosen next, then there are 9 choices of secretary. There remain eight people from whom to choose the treasurer, so the total number of ways of choosing a chairperson, secretary and treasurer (if they must be different people) is

$$10 \times 9 \times 8 = 720.$$

Activity 21 *Chairperson, secretary, treasurer and vice-chairman*

Suppose the club with ten members want a chairperson, secretary, treasurer and also a vice-chairperson. If they must all be different people, in how many different ways can they be chosen?

Generalising from these examples, we have the following result.

Number of ways of choosing an ordered sample from a set of objects

Suppose there are n objects to choose from. Then the number of ways of choosing x objects in a specified order is

$$n \times (n-1) \times \cdots \times (n-x+1).$$

(There are x terms in $n \times (n-1) \times \cdots \times (n-x+1)$.)

What we would actually like to count is the number of ways we can select, say, three people from ten *when the order in which the people are selected does not matter*. For example, if a committee of three people is to be formed and they will not have individual roles, then choosing A, then C and then D as the committee would be the same as choosing C, then A and then D.

So, how many ways can we form a committee of three people from ten members?

To tackle this question we can think of the task of choosing a chairperson, secretary and treasurer as a two-stage procedure: first we select a committee of three people, and then we allocate those three people the roles of chairperson/secretary/treasurer. Now the number of ways of allocating three roles to three people is $3 \times 2 \times 1$, as there are three choices of which committee member to make chairperson, then two committee members left for choosing the secretary, and then the one remaining committee member becomes treasurer. Thus, for every choice of committee, there are $3 \times 2 \times 1$ ways of choosing a chairperson, secretary and treasurer. Consequently,

number of ways of choosing a chairperson, secretary and treasurer
$= 3 \times 2 \times 1 \times$ (number of ways of choosing a committee of three people).

Also, from Example 6, we know that the total number of ways of choosing a chairperson, treasurer and secretary from ten people is $10 \times 9 \times 8$. Thus

$10 \times 9 \times 8$
$= 3 \times 2 \times 1 \times$ (number of ways of choosing a committee of three people).

So the number of ways of choosing a committee of three people from ten members is

$$\frac{10 \times 9 \times 8}{3 \times 2 \times 1} = 120.$$

Activity 22 *Committee of four people*

Suppose a club has 12 members.

(a) In how many ways can a chairperson, vice-chairperson, secretary and treasurer be chosen if no person can hold more than one role?

(b) Suppose we have selected four people. In how many ways can they be allocated the roles of chairperson, vice-chairperson, secretary and treasurer?

(c) Hence, what is the number of ways in which a committee of four people can be selected, if the order of selection does not matter?

If we make a selection, and the order in which the people or items are selected does not matter, then the selection is referred to as a **combination**. If there are n people, we let ${}^nC_x$ (read as 'n choose x') denote the number of ways of choosing a Combination of x people.

Thus ${}^{10}C_3$ is the number of ways of choosing, from 10 members, a committee of size 3. Similarly, ${}^{15}C_6$ is the number of ways of choosing, from 15 objects, a collection of 6 objects if the order in which they are chosen is of no importance. We have that

$$ {}^{10}C_3 = \frac{10 \times 9 \times 8}{3 \times 2 \times 1}. $$

Note that there are three terms in both the numerator and denominator of ${}^{10}C_3$. Also,

$$ {}^{15}C_6 = \frac{15 \times 14 \times 13 \times 12 \times 11 \times 10}{6 \times 5 \times 4 \times 3 \times 2 \times 1}, $$

where now there are six terms in both the numerator and denominator.

Factorial notation

You may have come across factorial notation in other modules and know that

$$ 3! = 3 \times 2 \times 1, $$
$$ 7! = 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1, \quad \text{and} $$
$$ 10! = 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1. $$

In factorial notation,

$$ {}^{10}C_3 = \frac{10!}{7!\,3!} $$

and it would often be written in this way. However, factorial notation will not be used in this module.

The following summarises the results we have obtained so far in this subsection.

Number of ways of choosing an unordered sample from a set of objects

Suppose there are n objects to choose from. Then the number of ways of choosing x objects if the order does not matter is

$$ {}^{n}C_x = \frac{\text{number of choices of } x \text{ objects if order does matter}}{\text{number of ways in which } x \text{ objects can be ordered}} $$
$$ = \frac{n \times (n-1) \times \cdots \times (n-x+1)}{x \times (x-1) \times \cdots \times 1}. $$

(There are x terms in both $n \times (n-1) \times \cdots \times (n-x+1)$ and $x \times (x-1) \times \cdots \times 1$.)

For any value of n, ${}^{n}C_0 = 1$ and ${}^{n}C_n = 1$. This can be seen directly from the definition, because there is only one way that zero objects can be selected from n objects and only one way that n objects can be selected from n objects.

Activity 23 *Combinations*

Calculate ${}^8C_3$, ${}^7C_5$ and ${}^4C_1$.

More combinations? (of clothes, that is)

3.2 Probabilities of combinations

Here we again suppose that we have a sample of observations from a population. How many of these observations are greater than the population median will partly be a matter of chance. In this section we aim to determine the probability that the number will be 1, the probability that it will be 2, the probability it will be 3, and so forth. This will answer the question posed at the start of Section 3:

> *If we have a sample of size 12, what is the probability that exactly nine values will exceed the population median?*

When the size of the sample exceeds 5 or 6, we will need our results about counting combinations in order to determine the probabilities we want. We shall start, though, by considering the simplest possible case: a random sample of size 1. What is the probability that the selected value lies above the population median? We shall write this probability as $P([+])$, and we shall write $P([-])$ for the probability that the selected value lies below the population median. Using the definition from Section 2, we obtain

$$P([+]) = \frac{\text{number in population above median}}{\text{total number in population}}.$$

The definition of the population median is that half the population lies above it and half below it, so

$$P([+]) = \tfrac{1}{2}.$$

Similarly, $P([-]) = \frac{1}{2}$.

We are ignoring the possibility that the value is exactly equal to the population median.

Since the selected value must lie either above or below the median, one of these events must happen; that is, it must be the case that

$$P([+]) + P([-]) = 1.$$

Our calculated probabilities agree with this $\left(\frac{1}{2} + \frac{1}{2} = 1\right)$.

Moving on to a sample of size 2, let $([+],[-])$ mean that the first observation was $[+]$ and the second was $[-]$. Then

$$P([-],[-]) = P([-]) \times P([-]) = \tfrac{1}{2} \times \tfrac{1}{2} = \left(\tfrac{1}{2}\right)^2 = \tfrac{1}{4},$$
$$P([-],[+]) = P([-]) \times P([+]) = \left(\tfrac{1}{2}\right)^2 = \tfrac{1}{4},$$
$$P([+],[-]) = P([+]) \times P([-]) = \left(\tfrac{1}{2}\right)^2 = \tfrac{1}{4},$$
$$P([+],[+]) = P([+]) \times P([+]) = \left(\tfrac{1}{2}\right)^2 = \tfrac{1}{4}.$$

If we let $P(x[+])$ denote the probability that the number of $[+]$ is x, then $P(0[+]) = P([-],[-]) = \frac{1}{4}$ and $P(2[+]) = P([+],[+]) = \frac{1}{4}$, while

$$P(1[+]) = P([-],[+]) + P([+],[-]) = \tfrac{1}{4} + \tfrac{1}{4} = \tfrac{1}{2}.$$

Exactly one of the events $0[+]$, $1[+]$ and $2[+]$ happens, so

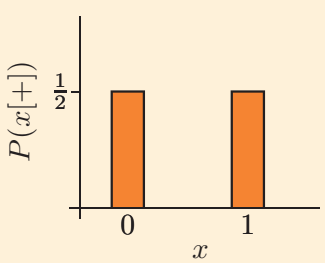
$$P(0[+]) + P(1[+]) + P(2[+]) = \tfrac{1}{4} + \tfrac{1}{2} + \tfrac{1}{4} = 1.$$

A **probability distribution** gives the probability of each possible event that could occur.

When the number of possible outcomes is small, a probability distribution is easily presented as a table or a bar chart. Figure 4 shows the probability distributions that were calculated above for a sample of size 1 and a sample of size 2.

Sample of size 1

Number of $[+]$s, x	0	1
$P(x[+])$	$\frac{1}{2}$	$\frac{1}{2}$



Sample of size 2

Number of $[+]$s, x	0	1	2
$P(x[+])$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

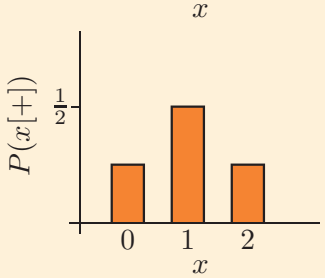


Figure 4 Probability distributions for samples of size 1 and size 2

Activity 24 *Probability distribution for a sample of size 3*

Suppose a sample of three observations is to be taken from a population.

(a) For each of the four numbers 0, 1, 2, 3, determine the probabilities that the number of observations greater than the median will be the same as it. Check that these four probabilities sum to 1.

(b) Present these values in a table and draw a bar chart of the probability distribution.

Calculations to obtain the probability distribution get longer as the sample size increases. Hence, for samples larger than 3 we need a better method for finding these probabilities. Let us consider the case where the sample size is 5.

Now the probability of the sequence [+], [−], [−], [+], [+] equals

$$\tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} = \left(\tfrac{1}{2}\right)^5.$$

Similarly, the probability of the sequence [−], [−], [−], [+], [−] equals

$$\tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} = \left(\tfrac{1}{2}\right)^5.$$

Indeed, the probability of any sequence that we specify will equal $\left(\frac{1}{2}\right)^5$. Consequently,

$$P(\text{number of [+]s equals } x) = (\text{number of sequences that have } x \text{ [+]s}) \times \left(\tfrac{1}{2}\right)^5.$$

Now to get a sequence that has x [+]s, we must just choose x of the five observations to be [+], leaving the remainder of them to be [−]. For example, if x is 2 and we label the five observations A, B, C, D and E, then the two [+] values could be chosen as A and C, for example, or A and E, or D and B, and so on. Hence, the number of sequences that have x [+]s is ${}^5C_x$, so

$$P(\text{Number of [+]s equals } x) = {}^5C_x \times \left(\tfrac{1}{2}\right)^5.$$

More generally, we have the following result.

Suppose we take a random sample of size n. Then

$${}^nC_x \times \left(\frac{1}{2}\right)^n$$

is the probability that exactly x of these observations are greater than the population median.

(The formula above is a special case of a probability distribution known as the 'binomial distribution'.)

You have now covered the material related to Screencast 4 for Unit 6 (see the M140 website).

Example 7 *Probability distribution for a sample of size 5*

A random sample of size 5 is selected.

1. What is the probability distribution for the number of values that lie above the population median?
2. What is the probability that exactly four of the selected values lie above the population median?
3. What is the probability that at least four values lie above the population median?

For the first question, applying the formula gives:

$$P(0[+]) = {}^5C_0 \times \left(\frac{1}{2}\right)^5 = \left(\frac{1}{2}\right)^5 = \frac{1}{32} = 0.031\,25,$$

$$P(1[+]) = {}^5C_1 \times \left(\frac{1}{2}\right)^5 = \frac{5}{1} \times \left(\frac{1}{2}\right)^5 = \frac{5}{32} = 0.156\,25,$$

$$P(2[+]) = {}^5C_2 \times \left(\frac{1}{2}\right)^5 = \frac{5 \times 4}{2 \times 1} \times \left(\frac{1}{2}\right)^5 = \frac{10}{32} = 0.3125,$$

$$P(3[+]) = {}^5C_3 \times \left(\frac{1}{2}\right)^5 = \frac{5 \times 4 \times 3}{3 \times 2 \times 1} \times \left(\frac{1}{2}\right)^5 = \frac{10}{32} = 0.3125,$$

$$P(4[+]) = {}^5C_4 \times \left(\frac{1}{2}\right)^5 = \frac{5 \times 4 \times 3 \times 2}{4 \times 3 \times 2 \times 1} \times \left(\frac{1}{2}\right)^5 = \frac{5}{32} = 0.156\,25,$$

$$P(5[+]) = {}^5C_5 \times \left(\frac{1}{2}\right)^5 = \frac{5 \times 4 \times 3 \times 2 \times 1}{5 \times 4 \times 3 \times 2 \times 1} \times \left(\frac{1}{2}\right)^5 = \frac{1}{32} = 0.031\,25.$$

These form the probability distribution, and the following table displays it in a clear form.

Table 7 Probability distribution for a random sample of size 5

Number of [+]s, x	0	1	2	3	4	5
$P(x[+])$	$\frac{1}{32}$	$\frac{5}{32}$	$\frac{10}{32}$	$\frac{10}{32}$	$\frac{5}{32}$	$\frac{1}{32}$

For the second question, we can now read off the first probability we require:

$$P(4[+]) = \frac{5}{32}.$$

The answer is quite satisfactory in this form. There is no need to write it as a decimal. If you do, it is reasonable to round it to three decimal places and write $P(4[+]) \simeq 0.156$.

The third question asks for the probability that at least four values in the sample lie above the population median. The only possibilities are four values or five values. As these are mutually exclusive events, we use the addition rule:

$$
\begin{aligned}
P(\text{at least 4[+]s}) &= P(4[+] \text{ or } 5[+]) \\
&= P(4[+]) + P(5[+]) \\
&= \frac{5}{32} + \frac{1}{32} \qquad \text{(from Table 7)} \\
&= \frac{6}{32} \\
&\simeq 0.188.
\end{aligned}
$$

Activity 25 *Probability distribution for a sample of size 4*

A random sample of size 4 is selected.

(a) Calculate the probability distribution for the number of values that lie above the population median. Check that the probabilities are all positive and add to 1.

(b) What is the probability that:

- two of the selected values lie above the population median and two lie below it?
- all the selected values lie on the same side of the population median?

In the question posed at the end of Section 1, we looked at data from a random sample of 12 large schools in the East of England and asked the question:

> *How likely is it that a random sample of 12 would contain nine values above and three values below the assumed population median?*

We repeated the question at the start of Section 3. You are now able to answer it!

Activity 26 *Probability that three out of 12 are below the median*

Suppose we take a random sample of size 12. What is the probability that there are exactly 9 values above the median?

Later we will need the complete probability distribution for the number of values that exceed the population median in a random sample of size 12. Determining this would be tedious using a calculator, so a computer would always be used. The probability distribution that would be obtained is given in the form of a diagram in Figure 5. The probabilities in the figure are rounded to three decimal places.

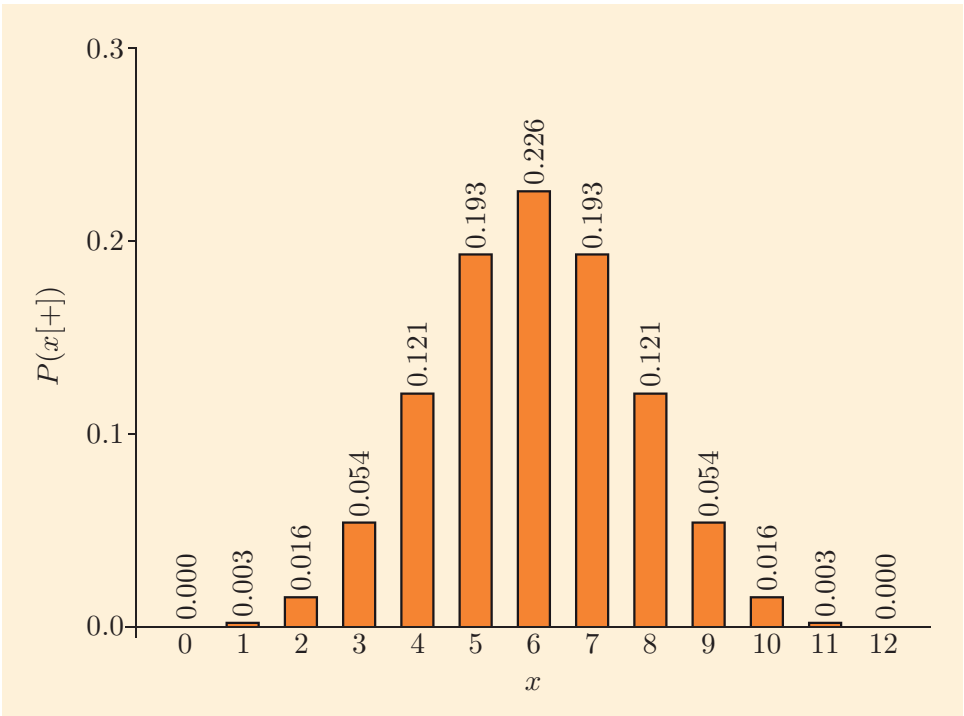


Figure 5 Probability distribution for a random sample of size 12

We can read off from this diagram the probability of any specific outcome. In Section 4 we shall see how to use this information to make an inference about the population from the sample result.

Exercises on Section 3

Exercise 4 *Flag signals*

A box contains seven flags, each of a different colour. A signal is made by flying three of the flags on a vertical flagpole.

(a) How many different signals can be made? (Note that a signal depends on the order of the flags on the flagpole as well as their colour.)

(b) When a signal is flying, how many different combinations of flags could be left in the box?

Exercise 5 *Tossing a coin five times*

What is the probability that you get exactly three heads when you toss a coin five times?

4 Testing hypotheses I

In Subsection 1.2 we introduced a measure of truancy: a pupil's truancy rate is the proportion of days that the pupil was absent without authorisation, and a school's truancy rate is the average truancy rate of its pupils. The median truancy rate for all secondary schools in the East of England was 0.98%, and we wondered how the median truancy rate for *large* secondary schools in the East of England compared with this.

In a random sample of 12 large (secondary) schools in the East of England, 9 had a truancy rate above 0.98%. This is more than you would expect if the median for large schools were truly 0.98% – in a sample of 12 you would get 6 values above the median, on average. The question is:

> *Given these sample data, could we reasonably believe that the truancy rate in large schools in the East of England is 0.98%?*

The procedure of using probability to examine this type of question – where sample data are used to evaluate the credibility of a statement – is called a **hypothesis test**. We will first give the steps in a hypothesis test and then discuss them.

Steps in a hypothesis test

1. Make a statement about the population of interest (e.g. *the median truancy rate of large schools in the East of England is 0.98%*). This is the hypothesis we wish to test.
2. Under the assumption that the hypothesis is true, determine the probability distribution for all possible values of some sample statistic (e.g. determine the probability distribution for the number of large schools, out of 12, that will have a truancy rate above 0.98%).
3. Now take the sample and ascertain *how unlikely* the observed value of the sample statistic is, on the basis of (1) and (2) (e.g. are we very unlikely to get a sample statistic as large as 9, that is, a sample of 12 schools in which nine or more schools have a truancy rate above 0.98%?).
4. If the sample turns out to have a very unlikely value, then either:
   - a very unusual event has happened, or
   - the hypothesis suggested in step 1 is incorrect, in which case the sample has provided evidence, albeit in a negative way, that adds something to our beliefs about the population.

Usually the hypothesis stated in step 1 is something that we do not really believe and would like to disprove. This is similar to a trial in a law court, and, indeed, hypothesis tests have much in common with a law-trial:

A person is presumed innocent until proven guilty.

- In a law-trial, the hypothesis is that the person in the dock is innocent. The prosecution do not really believe this, though, which is why the person is in the dock.
- The evidence is examined: a witness identified the defendant as being at the scene of the crime; the defendant has no alibi for the time of the crime; a footprint matches one of the defendant's shoes; etc. The fundamental question is: how likely are these occurrences if the defendant is innocent?
- When the events are extremely unlikely to have arisen *if the defendant is innocent*, it is concluded that the assumption of innocence is wrong, and the defendant is found 'guilty'; i.e. it is concluded that, beyond all reasonable doubt, the hypothesis of innocence is wrong.

There have been cases where probabilities could be calculated under the assumption of a defendant's innocence. This raises the question: in a law trial, when is a probability sufficiently small to equate to 'beyond all reasonable doubt'?

4.1 Tackling the problem

Much of the content in Units 6 to 8 is concerned with making a hypothesis about a population, and then analysing samples to test whether the hypothesis is reasonable. We shall meet several different hypothesis tests; the one we introduce here is called the **sign test** because it is based on the number of [+]s and [−]s.

Before giving a formal test, the following activity asks you to make your own intuitive judgement about whether sample data indicate that a hypothesis is wrong. The hypothesis is that the median truancy rate of large schools in the East of England is 0.98. The data are the truancy rates in a random sample of 12 such schools.

Activity 27 *What is your judgement?*

For each of the following instances of random samples of 12 schools, would you think the hypothesis 'The median truancy rate of large schools in the East of England is 0.98' is (i) quite possibly true, (ii) probably wrong, or (iii) almost certainly wrong?

(a) All 12 values are above the hypothesised median.

(b) There is 1 value above and 11 values below the hypothesised median.

(c) There are 6 values above and 6 values below the hypothesised median.

(d) There are 7 values above and 5 values below the hypothesised median.

(e) There are 9 values above and 3 values below the hypothesised median.

The solution to Activity 27 suggests that if all, or nearly all, the values lie on the same side of the assumed median, as in (a) and (b), then we should conclude that the hypothesis is almost certainly wrong. If the values are fairly evenly distributed on both sides of the assumed median, as in (c) and (d), then we should conclude that the hypothesis is quite possibly true. When the outcome is between these two, as in (e), then we are not sure.

Clearly sample data can cast doubt on the truth of a hypothesis. But how much doubt? Just using intuition to judge this is unsatisfactory. Given sample data, we need an objective way of quantifying the evidence that a hypothesis is wrong. In hypothesis testing, we use the probability distribution of all possible outcomes to measure the evidence that a hypothesis is wrong.

Figure 5 gave the probability of each possible outcome when we take a sample of size 12 and the probability of [+] for each item is $\frac{1}{2}$. Figure 5 is reproduced here as Figure 6.

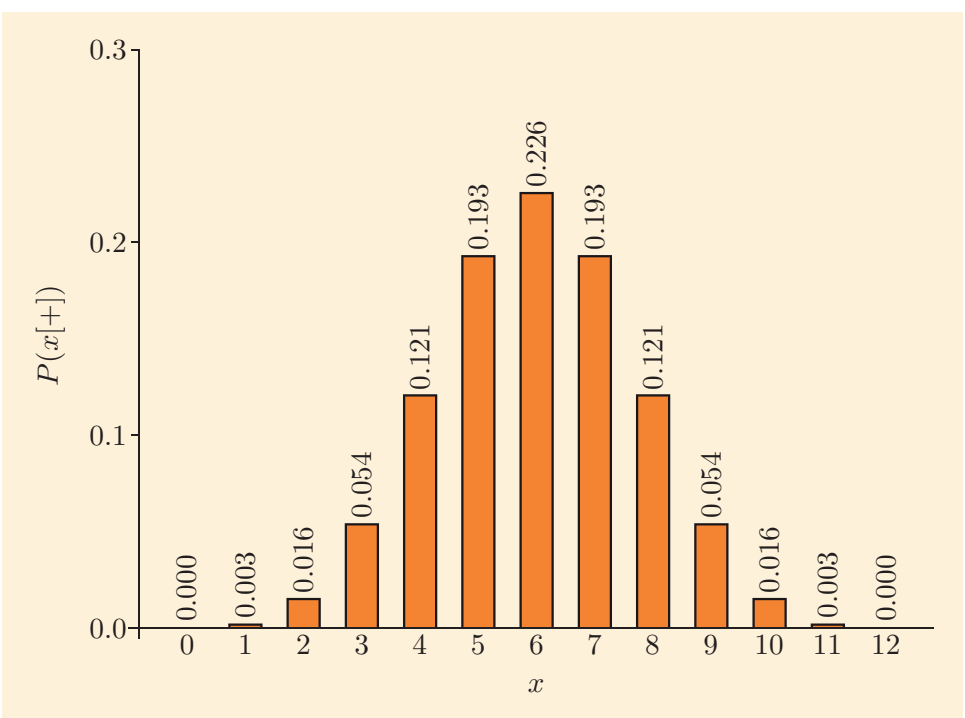


Figure 6 Probability distribution for a random sample of size 12

In a court of law, an unequivocal decision must be made as to whether or not a defendant is guilty. Here, for the moment, we shall assume that we must also decide one way or the other on whether or not to reject the hypothesis about the median. Thus we want a rule – a way of deciding – that tells us how extreme the outcome has to be for us to reject the hypothesis. In Figure 6 you can see that all the individual probabilities are quite small; even the probability of getting 6[+]s and 6[−]s is only 0.226 (about $\frac{2}{9}$, so between 1-in-4 and 1-in-5 chance). However, the extreme probabilities are much smaller than this. $P(0[+])$ and $P(12[+])$ are so small they appear as 0.000.

So *if* the hypothesis is true, then the outcome 0[+] is extremely unlikely to occur, as is the outcome 12[+]. This supports our intuitive feeling that if we did select a random sample with either 0[+] or 12[+], then we should conclude that the hypothesis is almost certainly wrong. So one possible rule would be to reject the hypothesis if either of these extreme outcomes occurs. However, this rule is erring on the cautious side; it does not cover (b) in Activity 27, which is also very extreme. Another possible rule would be to reject the hypothesis if the sample contains 0[+], 1[+], 11[+] or 12[+].

Activity 28 *Probabilistic judgement*

(a) Use the probability distribution of Figure 6 to find the probability that a random sample of size 12 has one of the outcomes 0[+], 1[+], 11[+], 12[+].

(b) On the basis of your result in part (a), do you think it is reasonable to reject the hypothesis if a sample is selected with one of these outcomes?

The probability of getting one of these four extreme outcomes is only 0.006 – that is, 6 chances in 1000 or about 1 in 170. Since this probability is also very small, perhaps we are still being over-cautious if we decide to reject the hypothesis only if one of these four outcomes occurs. It might be sensible to incorporate the next most extreme outcomes, 2[+] and 10[+], into our rejection rule, and possibly also 3[+] and 9[+].

When you first come across the idea, you might think it sensible to reject the hypothesis *only* if there is a very small probability of the observed result. For example, you might think it a good idea to reject if all 12 observations are on the same side of the median. However, this means that we would not reject the hypothesis if 11 observations were above the median and only one below it. In such a case, it is much more likely that the hypothesis is wrong and the median is in fact larger than 0.98%. It is a matter of compromise; if we are over-cautious, we will often not reject a hypothesis which is false. We have to remember that we are working with probabilities when we look at samples. *We cannot draw conclusions with certainty; there will always be some doubt.*

How far should we continue? We shall go on adding extreme values into our rejection rule until their combined probability is too large for us to feel

justified in rejecting the hypothesis. What is an acceptable value for this probability? This depends on many factors like the importance of the decision, and we shall discuss this in later units. For the present, we shall choose a probability of 0.05 or 5%. This value is used very frequently in many practical situations, and is universally accepted as a reasonable choice.

Significance levels

We refer to 5% as the **significance level** of our hypothesis test.

If a sample is selected whose values are one of the 5% most extreme outcomes that might occur if the hypothesis were true, then *we reject the hypothesis at the 5% significance level.*

Note that this statement does not say 'we reject the hypothesis with absolute certainty'. Rather, the statement gives the probability that rejecting the hypothesis would be the wrong decision.

Use of the 5% significance level is illustrated in Figure 7; this is a general picture for any size of sample. The actual probabilities will be different for different sample sizes, but the general pattern is the same. The horizontal axis represents the number x of [+]s in the sample, and the vertical axis represents the probability of obtaining different values of x. The 5% most extreme values are shaded.

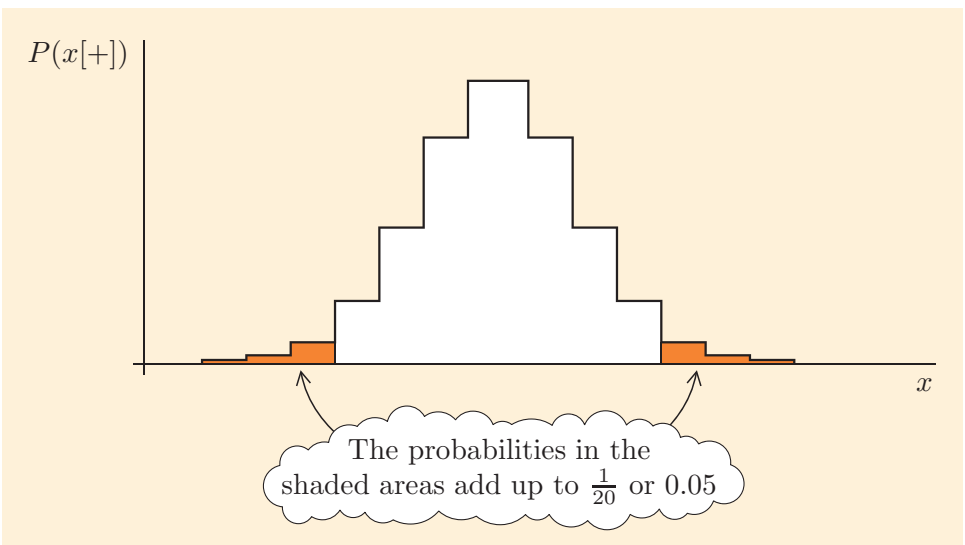


Figure 7 Tail areas adding to 5%

We shall now apply the rule to our random sample of size 12. First, notice that the probability distribution is symmetric, so we can start by considering the left-hand side only, as indicated by the shading in Figure 8. This means finding the outcomes in the left-hand tail whose probabilities add up to $\frac{1}{40} = 0.025$ or $2\frac{1}{2}\%$. Because the distribution is symmetric, there will be a corresponding 0.025 of probability in the right-hand tail, and the

two taken together will add up to the required 0.05 or 5%, as shown in Figure 8.

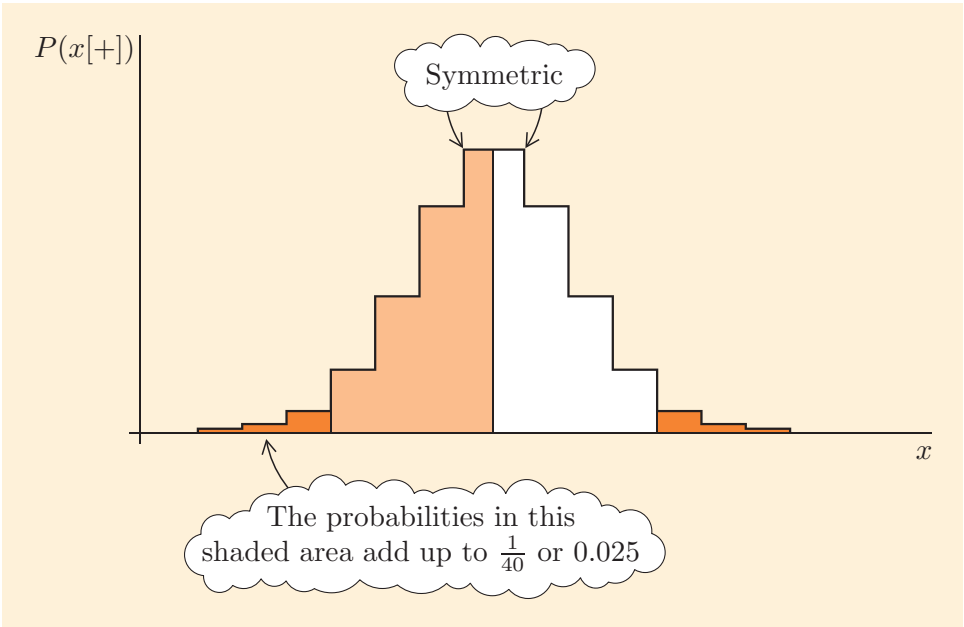


Figure 8 Symmetric distribution

Using the probability distribution of Figure 6 and the addition rule, we find the following probabilities for a random sample of size 12:

$$P(0[+]) \simeq 0.000,$$
$$P(0[+] \text{ or } 1[+]) = P(0[+]) + P(1[+]) \simeq 0.000 + 0.003 = 0.003,$$
$$P(0[+] \text{ or } 1[+] \text{ or } 2[+]) \simeq 0.003 + 0.016 = 0.019,$$
$$P(0[+] \text{ or } 1[+] \text{ or } 2[+] \text{ or } 3[+]) \simeq 0.019 + 0.054 = 0.073.$$

Notice that we only have to add the next probability to the previous total.

The probability of 0, 1 or 2 [+]s is 0.019 which is less than 0.025, whereas if we include 3[+]s the probability is greater than 0.025; neither is exactly equal to 0.025. We shall err on the side of caution and take the value which is less than 0.025. So we shall reject the hypothesis if a sample is selected with one of the outcomes 0[+], 1[+] or 2[+].

We shall also reject the hypothesis at the 5% significance level if the outcome is one of the corresponding extremes on the right-hand side, that is 10[+], 11[+] or 12[+]. To emphasise the symmetry, it is easier to think of these outcomes as 2[−], 1[−] and 0[−]. So for a sample of size 12, we shall reject the hypothesis if the number of values on one side of the assumed median is 2 or fewer. If our hypothesis is true, the probability that there are two or fewer values on one side of the median is $2 \times 0.019 = 0.038$. The probability that there are three or fewer values on one side of the median is $2 \times 0.073 = 0.146$ (about $\frac{1}{7}$). If we decided to reject the hypothesis when we observed three or fewer values on one side of the median, we would have a one in seven chance of rejecting the hypothesis when it was correct. Most people consider this an unacceptably high chance, and so we decide to reject the hypothesis if two or fewer values are on one side of the median.

The number 2 is called the **critical value at the 5% significance level** for a sample of size 12. We can use it to test a hypothesis that any random sample of size 12 comes from a population with some assumed median M. We simply count the number of values above and below M and check whether the smaller number of values is 2 or fewer; if it is, we reject the hypothesis.

This procedure does not apply only to random samples of size 12; for other sample sizes, we proceed in just the same way, though the critical value is different. In the next activity, you will find the critical value for a sample of size 15.

Activity 29 *Sample of 15 schools*

Suppose that we want to test the hypothesis that the population median is equal to 0.98%, and that we have a random sample of size 15 available. The probability distribution for this situation is shown in Figure 9.

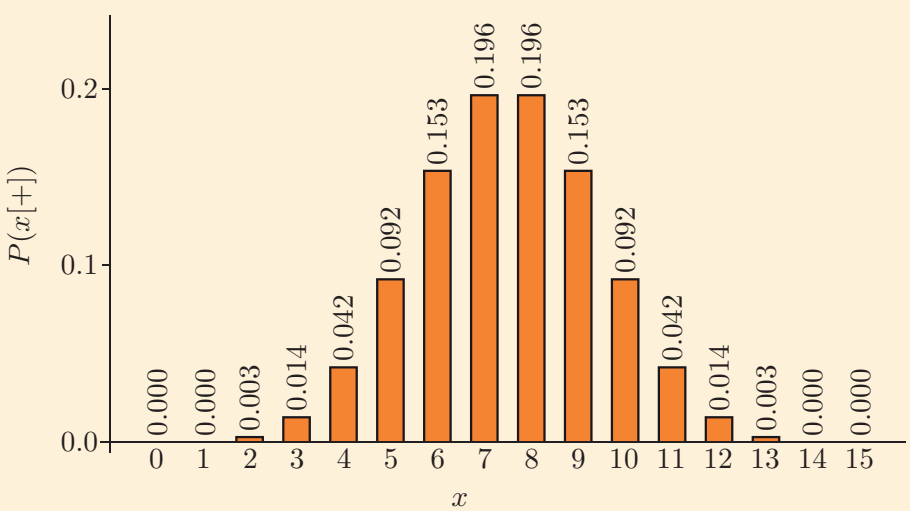


Figure 9 Probability distribution for a random sample of size 15

(a) Calculate the following probabilities:

- $P(0[+]) + P(1[+])$
- $P(0[+]) + P(1[+]) + P(2[+])$
- $P(0[+]) + P(1[+]) + P(2[+]) + P(3[+])$
- $P(0[+]) + P(1[+]) + P(2[+]) + P(3[+]) + P(4[+])$

(b) Which of your answers in part (a) is the largest value below 0.025?

(c) Write down the 5% most extreme outcomes, and state the critical value at the 5% significance level.

(d) Would you reject the hypothesis at the 5% level if the sample contained 14 values below the assumed median and 1 above it?

(e) Would you reject the hypothesis at the 5% level if the sample contained 5 values below the assumed median and 10 above it?

(f) Would you reject the hypothesis at the 5% level if the sample contained 12 values below the assumed median and 3 above it?

In Activity 29 you found that the critical value at the 5% significance level for a random sample of size 15 is equal to 3. In the same way, you could calculate the critical value for any size of sample, though you are not expected to do that in this module.

A typical probability distribution is shown in Figure 10. The shaded region corresponds to the outcomes

$$0[+], 1[+], \ldots, C[+]$$

and

$$C[-], \ldots, 1[-], 0[-],$$

where C is the critical value at the 5% significance level. The shaded area is called the **critical region at the 5% significance level**.

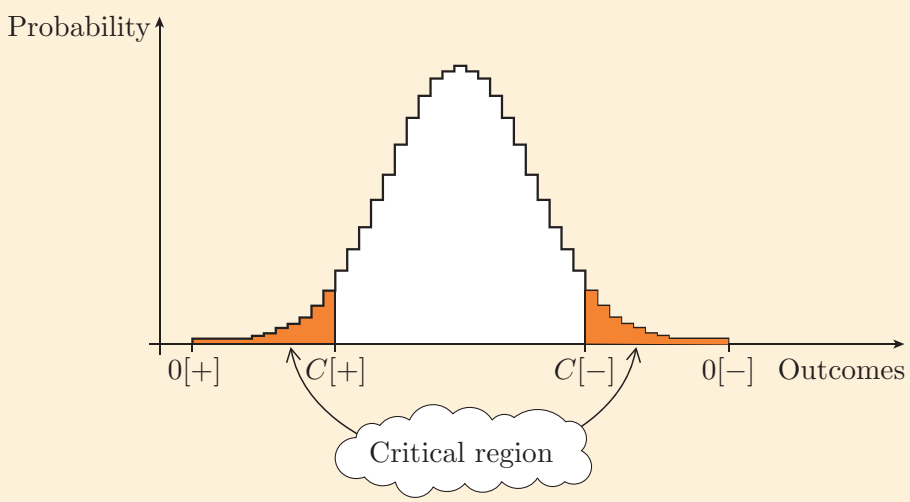


Figure 10 Critical region

Critical region and critical value

The critical *region* at the 5% level is chosen so that the combined probabilities of the outcomes falling in the region is 0.05 or just less than that.

The critical *value* at the 5% significance level can be used to determine whether or not an outcome is in the critical region.

For a sample of size 15, the critical value is $C = 3$, and the critical region contains the outcomes 0[+], 1[+], 2[+], 3[+], 0[−], 1[−], 2[−] and 3[−]. From Activity 29, it follows that the combined probability of these outcomes is $2 \times 0.017 = 0.034$, which is less than 0.05.

Remember that for a sample of size 15, 2[+] and 13[−] describe the same outcome.

The procedure can be used to calculate critical values and critical regions for samples of any size. Figure 10 illustrates the critical region for a moderately large sample.

The critical values at the 5% significance level have been calculated for random samples of sizes 1 to 40 and are given in Table 8. (The table could be extended to sample sizes greater than 40, but that is unnecessary for this module.)

Table 8 Critical values at the 5% significance level

Sample size	Critical value at the 5% significance level	Sample size	Critical value at the 5% significance level
1	–	21	5
2	–	22	5
3	–	23	6
4	–	24	6
5	–	25	7
6	0	26	7
7	0	27	7
8	0	28	8
9	1	29	8
10	1	30	9
11	1	31	9
12	2	32	9
13	2	33	10
14	2	34	10
15	3	35	11
16	3	36	11
17	4	37	12
18	4	38	12
19	4	39	12
20	5	40	13

For the sample sizes 1, 2, 3, 4 and 5, no critical value is given. To see why this is the case, consider a sample size of 4, for example. We found the appropriate probability distribution in Activity 25 (Subsection 3.2). The probability of 0[+] is $\frac{1}{16} = 0.0625$, the same as the probability of 4[+]. So the probability of the most extreme outcomes, 0[+] or 4[+], is $\frac{1}{8} = 0.125$, which is greater than 0.05. So no sample is extreme enough for us to be able to reject the hypothesis at the 5% significance level. The same situation applies for sample sizes 1, 2, 3 and 5.

Activity 30 *Critical values*

Use Table 8 to find the critical value at the 5% level for the following sample sizes. For each sample, specify the outcomes for which the hypothesis would be rejected at the 5% significance level.

(a) Sample size 21 (b) Sample size 32 (c) Sample size 7

So we can easily decide whether a given sample outcome should lead us to reject a hypothesis at the 5% significance level. We simply count up the number of [+]s and the number of [−]s in the sample. If the smaller of these two numbers is less than or equal to the critical value, then we reject the hypothesis.

4.2 The sign test

In Subsection 4.1, we introduced the hypothesis test known as the sign test. You will meet other hypothesis tests later in this module. In this subsection we describe the procedure for the sign test.

All the information we require from the sample is the smaller of the numbers of [+]s and [−]s it contains. This number is called the **test statistic**. If the test statistic is less than or equal to the critical value, then we reject the hypothesis at the 5% significance level.

Suppose we want to test whether some population has a median value of M. Then we select a random sample and apply the following procedure.

Procedure: the sign test

1. State the hypothesis that the population median is M.
2. Count the number of values in the sample that are larger than M (denoted by [+]s) and the number of values that are smaller than M (denoted by [−]s). The smaller of these two values is the test statistic.
3. Use Table 8 (at the end of Subsection 4.1) to write down the critical value at the 5% significance level corresponding to the size of the sample.
4. Compare the test statistic with the critical value. If it is less than or equal to the critical value, then the hypothesis is rejected at the 5% significance level. If the test statistic is greater than the critical value, then the hypothesis is not rejected.

The symbols $\leq$ and $>$ are often useful in such contexts.

These steps can be summarised by the flow chart given in Figure 11.

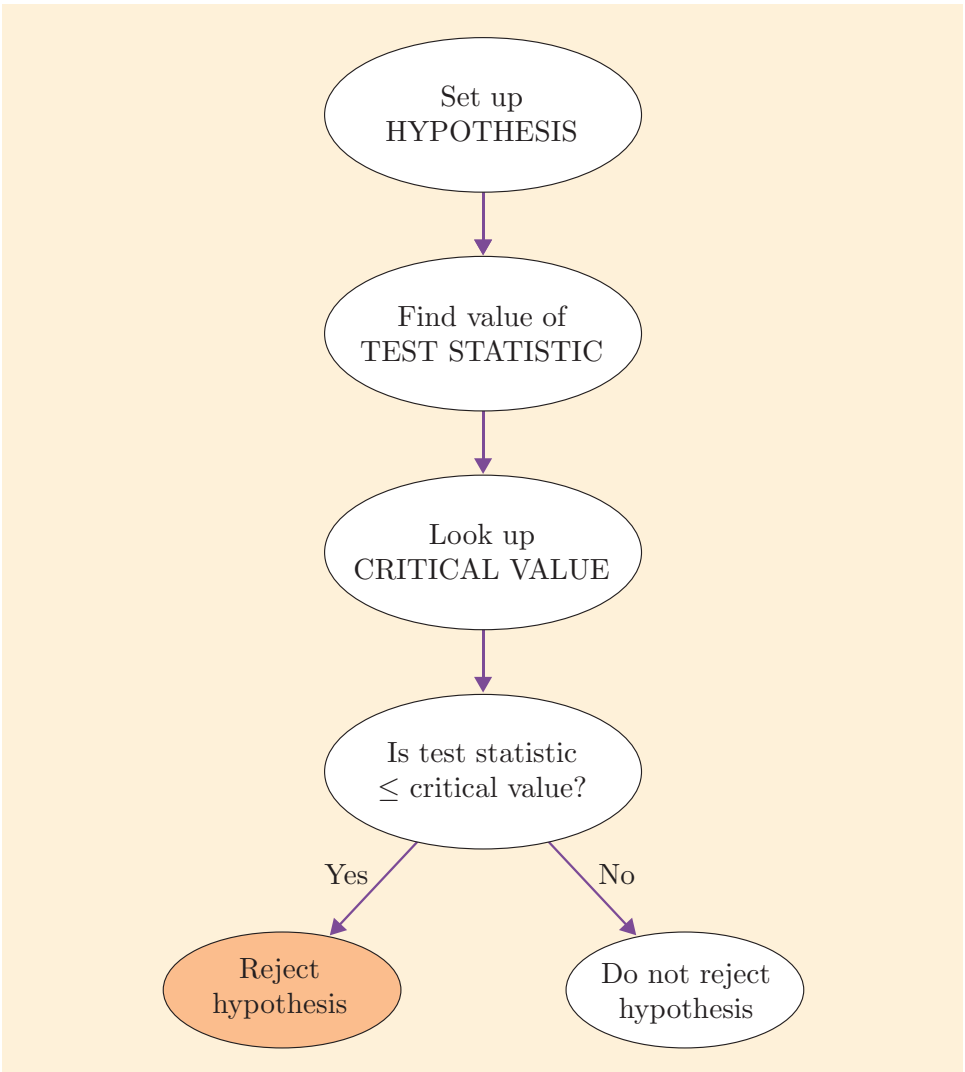


Figure 11 Steps in the sign test

To illustrate how to apply the sign test, let us complete the example we started in Section 1 and continued at the beginning of this section. We wanted to find out whether the truancy rate in large secondary schools in the East of England (those with over 1000 pupils) was the same as the overall rate for all secondary schools in the region. The truancy rate was recorded for a random sample of 12 large schools. We want to test whether this sample comes from a population whose median is 0.98%, the overall truancy rate for the East of England.

So we set up the hypothesis that the population median rate for large schools in the East of England is indeed 0.98%.

In the sample, we count the number of values above and below the assumed median of 0.98%. This gives 9 [+]s and 3 [−]s. The smaller of 9 and 3 is 3, so the test statistic is 3.

From Table 8, the critical value at the 5% significance level for a random sample of size 12 is equal to 2.

We then compare the test statistic with the critical value.

Since 3 is greater than 2, we cannot reject the hypothesis at the 5% significance level. This means that the sample we have observed could quite possibly have been drawn from a population with a median of 0.98%; there is insufficient evidence to suggest otherwise.

So hypothesis testing provides a method of inferring back from the sample to the population. However, it only enables us to reach a statistical conclusion, and there is more to interpreting results than this. We shall discuss this in Subsection 5.3, but for the moment you should concentrate on the statistical technique.

Activity 31 *Sample of 23 small schools*

In the example, we looked at data from large schools with over 1000 pupils. The data below show the truancy rates from a random sample of 23 small (secondary) schools (those with fewer than 500 pupils) in the East of England. You are asked to investigate whether the median percentage truancy rate for small schools is the same as the median percentage truancy rate for all schools in the East of England.

0.70	0.73	0.16	1.76	0.95	0.80	1.48	0.96	0.64	2.80	0.52	0.96
0.36	0.10	1.21	0.04	0.83	0.64	0.71	0.16	0.71	0.75	0.71	

(a) Write down the hypothesis to be tested.

(b) Record the number of values lying above and the number lying below the assumed median. Hence write down the test statistic.

(c) What is the appropriate critical value at the 5% significance level?

(d) Decide whether you would reject the hypothesis at the 5% significance level.

You have now covered the material related to Screencast 5 for Unit 6 (see the M140 website).

So far, we have applied the sign test to questions about truancy rates. It can be used in very many fields of application. One such application is shown next.

Activity 32 *Comparing two corn hybrids*

An experiment was conducted to compare two different hybrid lines of corn, Hybrid A and Hybrid B, with 28 plots used. Conditions were uniform within a plot, but differed between plots. To treat the hybrids comparably, each plot was divided in two: Hybrid A was grown on one half and Hybrid B on the other. The differences in yields from the two hybrids (Hybrid A − Hybrid B) are given in Table 9.

Examine whether one hybrid is better than the other by testing the hypothesis that the median difference between the two hybrids is 0.

Table 9 Differences in yield between two corn hybrids

1.7	−1.5	−0.6	−5.6	−1.3	−1.9	−10.3	−2.1	−0.6	0.6
−0.7	−0.5	0.1	0.7	0.9	−1.1	−0.5	−1.0	0.2	
−0.9	−0.2	−0.4	0.2	−0.7	−1.4	−0.4	−0.9	−1.5	

(Data source: Dixon, W. J. and Mood, A. M. (1946) 'The statistical sign test', *Journal of the American Statistical Association*, vol. 41, pp. 557–566)

Exercises on Section 4

Exercise 6 *Testing a claim about petrol consumption*

In Unit 1 (Section 3) we introduced values of petrol consumption for a Honda Civic 1.4i. The stemplot of these data is reproduced in Figure 12.

	26	1 7 9
	27	
	28	1 6
	29	6 7
	30	1
	31	7
	32	1 6
	33	5
	34	1
	35	0 2 2 3 3 7 8
	36	0 2 3 4 5 7
	37	5
	38	1 7 8
	39	1 3
	40	5
	41	
	42	1
$n = 34$	26	1 represents 26.1 miles per gallon

Figure 12 Stemplot of the petrol consumption data

Suppose that a dealer claimed that this type of car would give 37 miles to the gallon. Use the sign test to examine the truth of this claim.

Is this mouse one of the three?

Exercise 7 *Do mice like mirrors?*

An experiment was conducted to examine whether mice liked to have a mirror in their cage. There were 15 pairs of cages used. The cages in a pair were linked, and one cage in each pair had a mirror in it. A different mouse was placed in each pair, and the time that it spent in each cage was recorded. Three of the 15 mice spent more time in the cage with the mirror, while the other 12 mice spent more time in the cage without the mirror.

Test the hypothesis that the presence/absence of a mirror does not influence where a mouse spends its time.

5 Testing hypotheses II

In Section 4, you were introduced to hypothesis testing and the sign test in particular. In this section, you will learn more about hypothesis testing. We start with an approach that does not reduce the conclusions down to just 'reject' or 'do not reject'.

5.1 Significance probabilities: p-values

From Table 8 (Subsection 4.1), the critical value at the 5% significance level for a sample of size 20 is 5. Suppose, though, that in a random sample of 20 the number of values above the median (the number of [+] values) is only 3. Then, not only do we reject the hypothesised median at the 5% significance level, but we reject it very comfortably, as 3 is almost as close to 0 as it is to 5. Figure 13 gives the probability distribution for the number of values above the median in a random sample of size 20. From the figure, the probability of three or fewer [+] values is only $0.000 + 0.000 + 0.000 + 0.001 = 0.001$. Similarly the probability of three or fewer [−] values is 0.001. Hence, if we used a 0.2% significance level (rather than a 5% significance level), then the critical region would contain the outcomes 0[+], 1[+], 2[+], 3[+], 0[−], 1[−], 2[−] and 3[−], as the combined probability of these outcomes is $2 \times 0.001 = 0.002$. Stating that we reject a hypothesis at the 0.2% significance level is a much stronger statement than saying we reject it at the 5% significance level. Consequently, we would often like to be more precise when reporting the result of a hypothesis test, rather than simply saying how it compares with the 5% significance level.

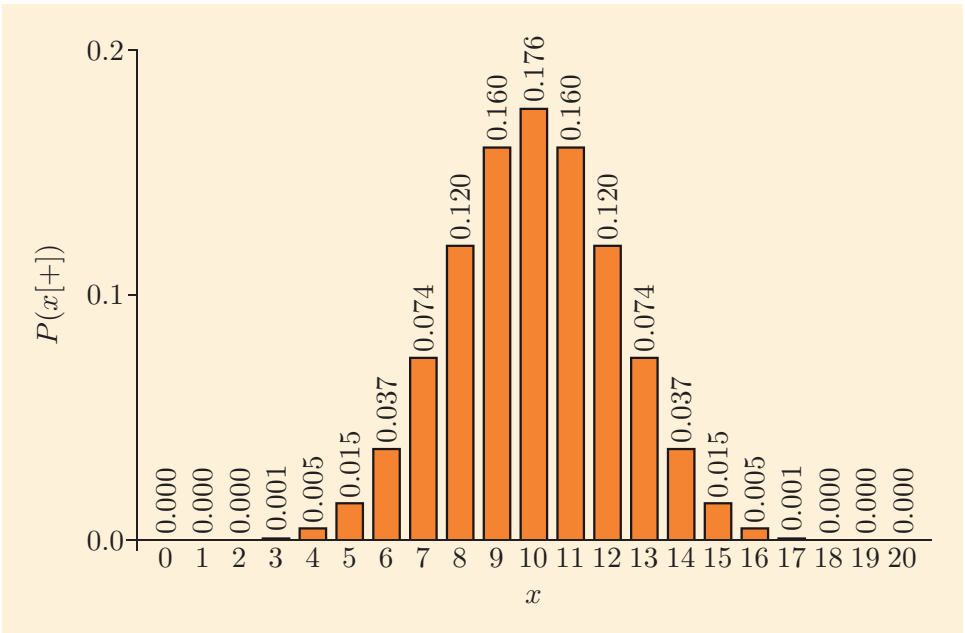


Figure 13 Probability distribution for a random sample of size 20

When we observe three [+] values in a sample of 20, we refer to 0.002 as the **significance probability** or, more commonly, as the ***p*-value** of the hypothesis test. If we set our significance level at anything larger than 0.2% (for example, we might set it to 5%), then we would reject the hypothesis, while if we set our significance level to anything smaller than 0.2%, then we would fail to reject the hypothesis.

Procedure: Obtaining p-values

To obtain the p-value (significance probability) for a hypothesis test, work through the following steps.

1. Assume the hypothesis is true.
2. Consider all the possible outcomes and divide these into two sets:

   Set A contains those outcomes that are as extreme or more extreme than the outcome that actually occurred.

   Set B contains those outcomes that are more likely than the outcome that actually occurred.
3. Calculate the probability that a random outcome would be from Set A. This probability is the p-value.

Set A is the smallest critical region that contains the outcome that actually occurred.

When data are analysed on a computer using a standard statistical package, the output almost always reports the result of a hypothesis test as a p-value.

A small p-value indicates that one of the more unlikely outcomes has occurred *or the hypothesis that led to the p-value is false.* As the p-value decreases, such an outcome becomes less likely and the evidence against

the hypothesis increases. Table 10 gives a reasonable way of interpreting different p-values. Notice that we never conclude that the hypothesis is true – a large p-value only means that there is little to suggest that the hypothesis is false.

Table 10 Interpretation of p-values

p-value	Rough interpretation
$p > 0.10$	Little evidence against the hypothesis
$0.10 \geq p > 0.05$	Weak evidence against the hypothesis
$0.05 \geq p > 0.01$	Moderate evidence against the hypothesis
$0.01 \geq p > 0.001$	Strong evidence against the hypothesis
$0.001 \geq p$	Very strong evidence against the hypothesis

At the end of a hypothesis test it is helpful to give a verbal description of the result and state the associated p-value in order to add precision. For example, we have focused on the hypothesis that the median truancy rate for large schools in the East of England is 0.98%. Using Figure 6, the p-value for the test when there are three [−] values out of 12 is

$$\begin{aligned} &P(0[-]) + P(1[-]) + P(2[-]) + P(3[-]) \\ &\quad + P(0[+]) + P(1[+]) + P(2[+]) + P(3[+]) \\ &= 2 \times [P(0[+]) + P(1[+]) + P(2[+]) + P(3[+])] \\ &= 2 \times (0.000 + 0.003 + 0.016 + 0.054) = 0.146. \end{aligned}$$

We might say, 'The p-value is 0.146, so there is little evidence against the hypothesis that the median truancy rate in large schools in the East of England is 0.98%.' Similarly, if there were only two [−] values out of 12, then the p-value would be $2 \times (0.000 + 0.003 + 0.016) = 0.038$, and we might say, 'The p-value is 0.038, so there is moderate evidence that the median truancy rate in large schools in the East of England is not 0.98%. The data suggest that the truancy rate is higher than that.'

You have now covered the material related to Screencast 6 for Unit 6 (see the M140 website).

Activity 33 *Essex schools: testing with p-values*

Essex is a county in the East of England region. To compare their truancy rates with the regional median of 0.98%, a random sample of 15 (secondary) schools in Essex was selected and their truancy rates recorded. These rates were as follows:

2.04 1.86 1.64 1.84 0.32 0.62 1.57 1.13
0.95 1.50 0.62 1.46 1.27 0.63 1.44

(a) What is the hypothesis to be tested?

(b) Record the number of values lying above the median, and the number lying below it. Hence write down the test statistic.

(c) Using Figure 9 (Subsection 4.1), calculate the p-value given by the hypothesis test.

(d) Making reference to this p-value, write down the conclusion to be drawn from the hypothesis test.

Activity 34 *Truancy rates in academies*

Suppose we took a random sample of secondary school academies in the East of England and examined their truancy rates. For each of the following sample results, test whether the median truancy rate for academies could be 0.98%.

Use Figures 6 and 9 (in Subsection 4.1) for the probabilities you need.

(a) The sample size is 12, with 11 values above 0.98% and 1 value below it.

(b) The sample size is 15, with 3 values above 0.98% and 12 values below it.

5.2 The sign test with ties

In the examples considered so far, there has never been a case in which one, or more, of the sample values is actually equal to the assumed population median. When this does happen, we shall refer to the situation as a **tie**. This problem is dealt with quite easily: we discard the tied values and reduce the size of the sample accordingly. The procedure is demonstrated in Example 8.

Some statisticians deal with ties in other ways. In this module we use the procedure described.

Example 8 *Testing with a tie*

In this unit, truancy rates have been re-calculated from government figures so as to determine them with an accuracy of two decimal places. In the government publication, the truancy rates are only given to an accuracy of one decimal place. Using the published rates, the median truancy rate for secondary schools in the East of England would be 1.0% (rather than 0.98%).

The data for the 15 schools in Essex (Activity 33) become:

2.0 1.9 1.6 1.8 0.3 0.6 1.6 1.1
1.0 1.5 0.6 1.5 1.3 0.6 1.4

Suppose we want to use these data to test whether the median truancy rate in Essex differs from a population whose median is 1.0%. There are 10 values above and four values below 1.0%, and one value actually equals 1.0%. We could write this as

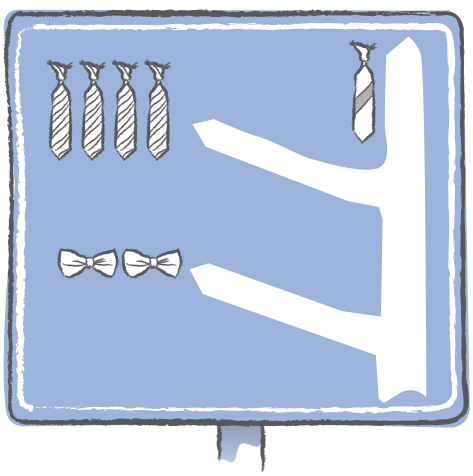
ten [+]s, four [−]s and one [=].

We discard the [=] and treat the sample as one of size 14 with ten [+]s and four [−]s. So the test statistic is 4. From Table 8 (Subsection 4.1), the critical value is 2. Since 4 is greater than 2, we cannot reject the hypothesis, and we conclude that it is quite possible that the median truancy rate in Essex is 1.0%.

In general, the procedure for dealing with ties is as follows.

Procedure: the sign test with ties

For a sample of size n containing m ties (that is, m of the sample values are equal to the assumed median), discard the m ties and treat the sample as one of size $(n - m)$.



Mark one answer

You are going shopping and need to buy neckties for all your male relatives. Which way should you go?

- First left ☐
- Second left ☐
- Straight on ☐
- Do a U-turn ☐

Another sign test with ties!

Activity 35 *Small schools with some tied data*

When truancy rates are only recorded to one decimal place, the following are the rates for 23 small schools in the East of England. (The data are from Activity 31.)

0.7	0.7	0.2	1.8	1.0	0.8	1.5	1.0	0.6	2.8	0.5	1.0
0.4	0.1	1.2	0.0	0.8	0.6	0.7	0.2	0.7	0.8	0.7	

We want to test the hypothesis that the truancy rate of small schools in the East of England is 1.0%.

(a) How many values are above the hypothesised median, how many are equal to it, and how many are below it?

(b) What is the test statistic, and what is the sample size that should be used in the hypothesis test?

(c) Test the hypothesis and state your conclusion.

Activity 36 *Change in Sheffield temperature?*

Meteorological records are available that give the average daily maximum temperature in Sheffield in August for each year from 1883, apart from three years (1918, 1919 and 1923). The median of these temperatures from 1883 to 2011 is 19.7 °C. The following stemplot gives the average daily maximum August temperatures in Sheffield in the 30 years from 1982 to 2011. If the climate has not changed since 1883, then these temperatures should be a random sample from all the average daily maximum annual temperatures in 1883–2011. That is, the median of these temperatures should be 19.7 °C.

Will the August temperatures in Sheffield become like those experienced in Florence?

	17	1
	17	
	18	4 4
	18	
	19	2
	19	5 5 6 6 7 7 7 9
	20	1 1 1
	20	7 9
	21	0 1 2 2 3
	21	5 9
	22	1 4
	22	5
	23	1 3
	23	
	24	4
$n = 30$	17	1 represents 17.1 °C

Figure 14 Stemplot of the average daily maximum August temperature in Sheffield (1982–2011)

(Data source: Met Office, historic station data)

We want to test the hypothesis that the values in the stemplot are a random sample from a population whose median is 19.7 °C.

(a) How many values are above the hypothesised median, how many are equal to it and how many are below it?

(b) What is the test statistic, and what is the sample size that should be used in the hypothesis test?

(c) Test the hypothesis and state your conclusion.

5.3 Conclusions and reservations

In Sections 4 and 5, the focus has been on truancy rates in different types of (secondary) schools in the East of England, where the rate is 0.98%. We found evidence that small schools in the region had a lower median truancy rate than 0.98%, while there was little evidence that large schools or schools in Essex had a different rate. However, in drawing a conclusion from a hypothesis test, or making any other form of statistical inference, we should think a bit about whether we have any reservations about the analysis. Let us do that now.

First, consider the data we used. Truancy rates for schools were probably obtained from class attendance registers. Presence or absence of each child is noted at the beginning of each morning and afternoon session. However, it is well known that some children attend for the register and then disappear either for the whole session or perhaps just for one particular lesson they dislike.

Also, some schools are probably more disciplined than others in collecting and recording the data. Indeed, because the data are published, it is conceivable that some schools might feel that a low truancy rate would attract new pupils. Other criticisms could also be raised, so we must have some reservations about the accuracy of our data. However, they are from the best publicly available source of information on truancy rates.

Second, let us think about the type of conclusion we drew. We decided that there was no reason to assume that the median truancy rate for large schools in the East of England was different from 0.98%, which was the overall rate for all schools in the East of England. Notice that we do not say that the rate for large schools is equal to 0.98%. The fact that we do not reject the hypothesis certainly does not justify us assuming that the median is equal to 0.98%. If we examined all the large schools in the East of England, we would almost certainly find that the median truancy rate for large schools differs in the East of England and is not exactly 0.98%. On the basis of the sample we have, though, it is unclear whether the median rate would turn out to be above or below 0.98%.

Note that a statistical test can tell us nothing about causation. For example, we concluded in Activity 31 that the truancy rate for small schools in the East of England is probably lower than the overall rate for all schools in the East of England. We cannot conclude from this that the small size of a school *causes* the truancy rate to fall. There may well be other factors, like the situation of the school, which influence both the size and the truancy rate.

The reservations expressed in this subsection may have left you feeling that it was not worth doing the statistical analyses. However, the analyses have added to our knowledge about truancy rates. The point is that a statistical analysis enables you to quantify the uncertainty that exists in drawing conclusions – in most other circumstances, the uncertainty of conclusions is either not recognised or is ignored.

'Thanks Mary, that does put our data in perspective.'

Exercises on Section 5

Exercise 8 *A second look at mice and mirrors*

In Exercise 7 (Section 4) an experiment was described in which 12 out of 15 mice preferred a cage without a mirror, while three mice preferred a cage with a mirror. The exercise asked for a test of the hypothesis that the presence/absence of a mirror does not influence where a mouse spends its time. Using the probabilities given in Figure 9 (Subsection 4.1), determine the p-value of the test and evaluate the evidence against the hypothesis.

Exercise 9 *Change in depression*

A psychologist rated 16 subjects undergoing withdrawal from narcotics on the basis of the extent of their depression before and one hour after receiving a dose of methadone. The degree of depression was rated on a scale from 1 (= no depression) to 5 (= severe depression). The results are given in Table 11.

Table 11 Depression scores before and after a dose of methadone, and their differences

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Before	2	1	2	3	2	2	2	2	2	1	4	3	2	3	3	2
After	1	1	3	1	1	2	4	1	1	1	2	2	1	1	3	3
Change	−1	0	1	−2	−1	0	2	−1	−1	0	−2	−1	−1	−2	0	1

If methadone does not affect depression, then the difference between a subject's 'before' and 'after' scores is equally like to be above or below 0. (It could also equal 0.) Thus, we want to test the hypothesis that the median difference between 'before' and 'after' scores is 0 – rejecting this hypothesis would show evidence that methadone affects depression.

(a) How many values are above the hypothesised median, how many are equal to it, and how many are below it?

(b) What is the test statistic, and what is the sample size that should be used in the hypothesis test?

(c) Test the hypothesis and state your conclusion.

---

6 Computer work: probabilities and the sign test

In this section you will use the statistical package Minitab to calculate probabilities of the type that you calculated (for small sample sizes) in Subsection 3.2. You will also learn how to use Minitab to perform a sign test for a population median. You should now turn to the Computer Book and work through Chapter 6.

Summary

In this unit, you have been introduced to the basic concepts of probability and hypothesis testing. These are incredibly important in statistics: probability underpins most statistical methods, and one of the most common tasks in statistical inference is hypothesis testing. A consequence is that the ideas introduced in this unit will be reinforced in subsequent units and, indeed, in any further statistics modules that you study.

You have learned the fundamentals of probability – its definition as a proportion, the addition rule and the multiplication rule. You have also learned how to count the number of ways that combinations can occur, so that you can now calculate the probability distribution for the number of sample data that will exceed the population median. You used this probability distribution to perform a sign test – while simultaneously learning the structure of a hypothesis test, the meaning of a p-value, and how to report your conclusions and consider reservations you might have about the test. The range of tasks you can perform with Minitab has been extended.

Learning outcomes

After working through this unit, you should be able to:

- appreciate how the modelling diagram from earlier units can be modified to take account of the important step of inferring back from the sample to the population
- appreciate that a general question needs to be clarified and made more precise before it can be answered using statistics
- calculate probabilities based on random selection
- apply the addition rule for mutually exclusive events
- apply the probability rule for complementary events
- apply the multiplication rule for statistically independent events
- count the number of ways that a specified combination can occur
- calculate the probability distribution for the number of values above the median of the population
- understand the concepts of a hypothesis test and of rejecting a hypothesis at the 5% significance level
- look up the critical value for the sign test at the 5% significance level, and use this value to determine the critical region
- apply the sign test, including the case when ties are present
- consider what reservations you have about the conclusion drawn from a hypothesis test
- use Minitab to calculate the probability distribution for the number of outcomes that exceed the median and perform a sign test.

Solutions to activities

Solution to Activity 1

There are many possible factors you could have thought of. Some are given here, but your list may be quite different and equally valid.

Work: children may absent themselves from a particular lesson because they dislike that subject, have not completed homework, etc.

Home background: parents who take their children to school and are interested in their education will encourage regular attendance.

Bullying: children may stay away because they are being bullied and are frightened to go to school.

Helping at home: children may be encouraged by their parents to miss school to help with housework or take care of young siblings, or they may feel they need to look after their parents.

Age: truancy is much commoner among older children approaching school-leaving age than among those in primary schools.

Solution to Activity 2

As with the previous activity, there are many possible answers here. Some are listed below.

Level: primary or secondary school

Location: inner city or rural school, for example

Size of school

Type of school: independent, academy, etc.

Solution to Activity 3

There are many possible reasons that are definitely truancy: missing school to go to the cinema/football match/fishing; missing school to avoid boredom/a lesson you hate; etc.

Reasons that would not normally be classed as truancy include illness, medical/dental appointments, attendance at family funerals, being kidnapped, religious observance.

Reasons that might (or might not) be truancy include staying off school because you didn't feel that well, when it would depend on how unwell you felt and how often it happened. Another example might be staying off school to look after a sick brother/sister/parent or go on holiday with parents, when it might again depend on how often this happened, and might also depend on whether the school had given permission.

Solution to Activity 4

In order to compare truancy in different schools, we must somehow take account of the number of pupils in each school. If we simply counted the number of cases of truancy, then large schools would almost inevitably have higher truancy counts. So we need some measure of the rate of truancy in a school. Here are three possible ways of collecting data and measuring truancy; you may well have thought of others.

1. Use the school attendance registers which record the children present at the beginning of each morning and afternoon. Record which of the children who were not present had notes of authorised absence due to illness or other reason and discount these cases. Calculate the proportion of children absent without authorisation, averaged over a year.
2. Choose a particular week. Count the number of children present for each lesson during that week, and divide this figure by the number of lessons and the number of children in the school. Subtract this number from 1 to give a truancy rate.
3. At some arbitrary time, send someone round the school to count the number of children present. Subtract this number from the total number on the school register to get the number who are absent. Divide by the total number of children to give a truancy rate.

None of these ways is ideal, but you can see that it is not a simple matter to obtain a reliable figure; there are many factors to consider.

Solution to Activity 5

(a) It would be possible to find the median of the 12 values by drawing a stemplot, but with so few, it is probably easier just to list them in order:

0.31, 0.82, 0.83, 1.06, 1.09, 1.19, 1.44, 1.52, 1.84, 1.88, 2.78, 2.90.

Since the sample size is even,

$$\begin{aligned}\text{median} &= \frac{\text{sum of two middle data values}}{2} \\ &= \frac{1.19 + 1.44}{2} \\ &= 1.315 \simeq 1.32.\end{aligned}$$

See Subsection 4.2 of Unit 1.

So the median of the sample is 1.32%.

(b) The median value of the truancy rate of the sample of 12 large East of England schools is somewhat larger than the median truancy rate for all the East of England schools. However, a different sample would probably have had a different median. In summary, the median truancy rate of all large schools in the East of England looks as if it might be larger than 0.98%, but we cannot be sure.

Solution to Activity 6

This is a suggested ranking of the events, from the most likely to the least likely.

C. The sun will rise tomorrow.

D. The sun will shine tomorrow.

F. You toss a coin and it lands tails up.

G. You throw a die and it shows a six.

B. Two out of a group of ten people have the same birthday.

A. Your new colleague at work has the same birthday as you.

E. You will win the jackpot in the National Lottery next week.

H. A member of a hockey team is 150 years old.

Your ranking may be slightly different from this, but it should be fairly similar.

(Note that event C can be regarded as certain to happen, whereas event H is impossible.)

Solution to Activity 7

(a) The probability of selecting a male student is

$$\frac{\text{number of male students}}{\text{total number of students}} = \frac{3293}{6082} \simeq 0.541.$$

(b) The probability of selecting a law student is

$$\frac{\text{number of law students}}{\text{total number of students}} = \frac{334}{6082} \simeq 0.055.$$

(c) The probability of selecting a female medical student is

$$\frac{\text{number of female medical students}}{\text{total number of students}} = \frac{206}{6082} \simeq 0.034.$$

Solution to Activity 8

(a) The probability of selecting a woman is

$$\frac{\text{number of women in population}}{\text{total number in population}}.$$

If we apply this definition to the case where there are 5 men and 5 women, then the probability of selecting a woman is

$$\frac{5}{10} = 0.5.$$

(b) For a population consisting of 1 man and 9 women, using the definition in the solution to (a) gives the probability of selecting a woman as

$$\frac{9}{10} = 0.9.$$

(c) For 9 men and 1 woman, the probability of selecting a woman is

$$\frac{1}{10} = 0.1.$$

(d) For 10 men and 0 women, the probability of selecting a woman is

$$\frac{0}{10} = 0.$$

(e) For 99 men and 1 woman, the probability of selecting a woman is

$$\frac{1}{100} = 0.01.$$

(f) For 0 men and 100 women, the probability of selecting a woman is

$$\frac{100}{100} = 1.$$

Solution to Activity 9

(a) For each of the events, the probabilities are as follows:

$$P(\text{child absent through truancy for 5 to 9 days}) = \frac{\text{total number absent for 5 to 9 days}}{\text{total number of children}} = \frac{90}{300} = 0.3.$$

$$P(\text{child absent for 5 to 9 days and attends School B}) = \frac{30}{300} = 0.1.$$

$$P(\text{child from School A is absent for 10 to 19 days}) = \frac{\text{number absent for 10 to 19 days from School A}}{\text{total number at School A}} = \frac{26}{200} = 0.13.$$

$$P(\text{child absent for} \geq 10 \text{ days}) = \frac{\text{number absent for 10 to 19 days} + \text{number absent for} \geq 20 \text{ days}}{\text{total number of children}} = \frac{45 + 15}{300} = \frac{60}{300} = 0.2.$$

(b) Let C be the event that a child is absent through truancy for 5 to 9 days and let B be the event that a child attends School B. (You might well have chosen different letters: that does not matter.)

Then

$$P(\text{child absent through truancy for 5 to 9 days}) = P(C)$$

and

$$P(\text{child absent for 5 to 9 days and attends School B}) = P(C \text{ and } B).$$

Solution to Activity 10

(a) The events are mutually exclusive, as a person can only have one blood type.

(b) The events are not mutually exclusive. Even though it is not very common, a person can have black hair and blue eyes.

(c) These are not mutually exclusive events. Although all three events cannot simultaneously happen, two of them could happen at the same time. (In fact, any two of them could happen at the same time.)

(d) These are not mutually exclusive events. While (i) and (ii) cannot occur simultaneously, either of them could occur with (iii).

Solution to Activity 11

(a) Altogether there are 200 children at School A. Using the data in Table 3, the probabilities can be calculated as follows.

There are 108 children absent through truancy for 0 to 4 days. Hence

$$P(\text{child from School A absent for 0 to 4 days}) = \frac{108}{200} = 0.54.$$

There are 60 children absent through truancy for 5 to 9 days. Hence

$$P(\text{child from School A absent for 5 to 9 days}) = \frac{60}{200} = 0.3.$$

Altogether $108 + 60 = 168$ children are absent through truancy for between 0 and 9 days. Hence

$$P(\text{child from School A absent for 0 to 9 days}) = \frac{168}{200} = 0.84.$$

(b) Let E denote the event that a child at School A is absent through truancy for 0 to 4 days, and F denote the event that the child is absent for 5 to 9 days. The events E and F are mutually exclusive, as both cannot apply to any one child.

From the solution to part (a), $P(E) = 0.54$ and $P(F) = 0.3$. The addition rule for mutually exclusive events states that

$$P(E \text{ or } F) = P(E) + P(F).$$

Now $P(E) + P(F) = 0.54 + 0.3 = 0.84$, which is the value we found in part (a) for the probability that a child at School A is absent through truancy for 0 to 9 days, that is $P(E \text{ or } F)$, so the addition rule holds in this case.

Solution to Activity 12

(a) Let M stand for the event that a man is selected and W stand for the event that a woman is selected.

$$P(M) = \frac{3293}{6082} \simeq 0.541.$$

Hence, $P(W) = 1 - P(M) \simeq 1 - 0.541 = 0.459$.

(b) $P(\text{selecting a medical student}) \simeq \dfrac{575}{6082} = 0.095.$

Hence

$$P(\text{selected student is not a medical student}) \simeq 1 - 0.095 = 0.905.$$

This is quicker than the alternative of putting:

$$P(\text{selected student is not a medical student}) = \frac{\text{number of students in science, arts and law}}{\text{number of students}}$$

$$= \frac{2602 + 2571 + 334}{6082} = \frac{5507}{6082} \simeq 0.905.$$

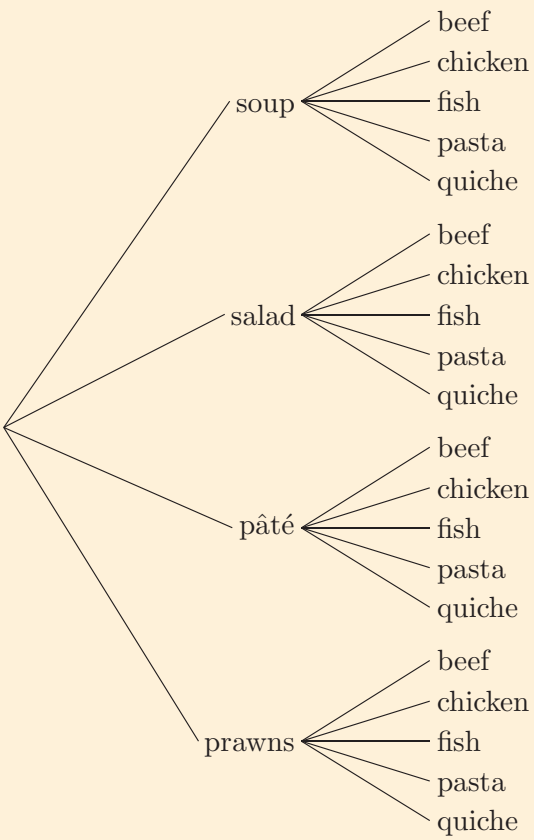
Solution to Activity 13

(a) 'G does not occur' is the event that a child selected at random from the two schools is absent through truancy for 0 to 4 days or is absent for 10 or more days.

(b) $$P(G \text{ does not occur}) = 1 - P(G)$$
$$= 1 - \frac{90}{300} = 0.7.$$

Solution to Activity 14

(a)



Tree diagram for a two-course meal

(b) There are four vegetarian combinations: soup–pasta, soup–quiche, salad–pasta and salad–quiche.

(c) $P(\text{vegetarian meal})$

$$= \frac{\text{number of vegetarian meal combinations}}{\text{total number of meal combinations}} = \frac{4}{20} = 0.2.$$

(d) $P(\text{vegetarian first course})$

$$= \frac{\text{number of vegetarian first courses}}{\text{total number of first courses}} = \frac{2}{4} = 0.5.$$

$P(\text{vegetarian second course})$

$$= \frac{\text{number of vegetarian second courses}}{\text{total number of second courses}} = \frac{2}{5} = 0.4.$$

The product of these answers is $0.5 \times 0.4 = 0.2$, which is the answer in (c).

Solution to Activity 15

(a) In part (a) of Activity 14, you found that there are 20 different two-course meal combinations. From each of these two-course combinations we can form three different three-course combinations, because there are three choices for the third course. Hence in total there are $20 \times 3 = 60$ different three-course combinations.

(b) $P(\text{third course is good with custard})$

$$= \frac{\text{number of third courses good with custard}}{\text{total number of third courses}} = \frac{2}{3} \simeq 0.667.$$

Two of the first courses are vegetarian, as are two of the second courses, and two of the third courses are good with custard. Hence the number of three-course meals that have two vegetarian courses followed by a course that is good with custard is $2 \times 2 \times 2 = 8$. (They are: soup–pasta–pie, soup–quiche–pie, salad–pasta–pie, salad–quiche–pie, soup–pasta–crumble, soup–quiche–crumble, salad–pasta–crumble and salad–quiche–crumble.) Hence

$P(\text{vegetarian first course } \textit{and} \text{ vegetarian second course } \textit{and} \text{ third course is good with custard})$

$$= \frac{8}{60} \simeq 0.133.$$

(c) $P(\text{vegetarian first course}) \times P(\text{vegetarian second course}) \times P(\text{third course is good with custard})$

$$\simeq 0.5 \times 0.4 \times 0.667 \simeq 0.133.$$

This is the same result as found in part (b) for

$P(\text{vegetarian first course } \textit{and} \text{ vegetarian second course})$.

Solution to Activity 16

(a) The events are not independent. Taller people tend to be heavier, so if a person is taller than average, then they are more likely to be heavier than average.

(b) These events are independent. The height of the first person has no influence on the weight of the second person, assuming the people were chosen at random. (If you choose two people standing next to each other, then the choices are not random, and the characteristics of one person could relate to the characteristics of the other.)

(c) These events are independent, as the day on which you are born has no influence on your weight.

(d) The events are not independent. For example, if the first two events both occur then we *know* that the third event occurs.

(e) The probability that the card is an ace is

$$\frac{\text{number of aces in pack}}{\text{number of cards in pack}} = \frac{4}{52} = \frac{1}{13}.$$

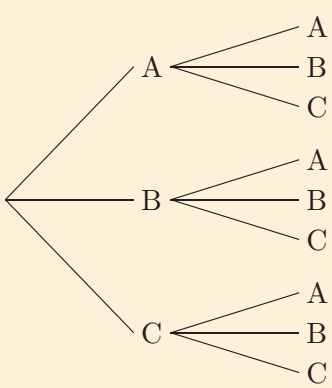
If we know that the card is a diamond, then the probability that it is an ace is

$$\frac{\text{number of cards that are the ace of diamonds}}{\text{number of diamonds}} = \frac{1}{13}.$$

These probabilities are the same. That is, knowing that one of the events occurs does not change the probability that the other event occurs. Hence the events are independent.

Solution to Activity 17

(a) Using A, B and C to represent Ashia, Brenda and Clare, the different possible samples are shown in the following tree diagram. So, the possible samples of size 2 are (A, A), (A, B), (A, C), (B, A), (B, B), (B, C), (C, A), (C, B) and (C, C). Altogether there are 9 possible samples.



(b) Because there are 3 possibilities for the first member of the sample and 3 possibilities for the second member, there are $3 \times 3 = 9$ possible samples altogether.

Solution to Activity 18

We have that $P(M) = 7/10 = 0.7$, so

$$P(2M) = P(M) \times P(M) = 0.7 \times 0.7 = 0.49.$$

Solution to Activity 19

At any selection we can write down the probabilities of selecting a man or a woman:

$$P(M) = \frac{40}{100} = 0.4 \quad \text{and} \quad P(W) = \frac{60}{100} = 0.6.$$

For a sample of size 2,

$$P(2M) = 0.4 \times 0.4 = 0.16,$$
$$P(0M) = P(2W) = 0.6 \times 0.6 = 0.36,$$
$$\begin{aligned} P(1M) &= P(\text{man selected first and woman selected second}) \\ &\quad + P(\text{woman selected first and man selected second}) \\ &= 0.4 \times 0.6 + 0.6 \times 0.4 \\ &= 0.24 + 0.24 = 0.48. \end{aligned}$$

As a check,

$$P(2M) + P(0M) + P(1M) = 0.16 + 0.36 + 0.48 = 1.00.$$

Solution to Activity 20

We are now considering only samples where the same person cannot be considered twice. There are three women, so the first one can be selected in 3 ways. Then only two women remain, so the second can be selected in 2 ways. Hence the total number of samples of size 2 consisting of two women is $3 \times 2 = 6$.

As a check, the samples of two women are (A, B), (A, C), (B, A), (B, C), (C, A), (C, B), where, for example, (A, B) means that Ashia is selected first and Brenda is selected second.

Solution to Activity 21

If they are picked in the order chairperson, secretary, treasurer, vice-chairperson, then there are 10 choices for chairperson, 9 choices for secretary, 8 choices for treasurer, and 7 choices for vice-chairperson, giving a total of

$$10 \times 9 \times 8 \times 7 = 5040 \text{ choices.}$$

Solution to Activity 22

(a) Here the order of selection matters. The chairman can be chosen from 12 members, the vice-chairman from the remaining 11, etc. The number of ways of choosing people for the 4 roles equals $12 \times 11 \times 10 \times 9 = 11\,880$.

(b) The number of ways of allocating 4 people the different roles is $4 \times 3 \times 2 \times 1 = 24$.

(c) The number of ways of choosing a committee of 4 people from 12 members is

$$\frac{\text{number of choices of 4 people if order of choice matters}}{\text{number of ways of allocating 4 people to 4 roles}} = \frac{11\,880}{24} = 495.$$

Solution to Activity 23

$$^8C_3 = \frac{8 \times 7 \times 6}{3 \times 2 \times 1} = \frac{336}{6} = 56.$$

$$^7C_5 = \frac{7 \times 6 \times 5 \times 4 \times 3}{5 \times 4 \times 3 \times 2 \times 1} = \frac{2520}{120} = 21.$$

$$^4C_1 = \frac{4}{1} = 4.$$

Solution to Activity 24

(a) The four probabilities required are:

$$P(0[+]) = P([-],[-],[-]) = P([-]) \times P([-]) \times P([-])$$
$$= \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} = \left(\tfrac{1}{2}\right)^3 = \tfrac{1}{8},$$

$$P(1[+]) = P([-],[-],[+]) + P([-],[+],[-]) + P([+],[-],[-])$$
$$= \left(\tfrac{1}{2}\right)^3 + \left(\tfrac{1}{2}\right)^3 + \left(\tfrac{1}{2}\right)^3 = 3 \times \left(\tfrac{1}{2}\right)^3 = \tfrac{3}{8},$$

$$P(2[+]) = P([-],[+],[+]) + P([+],[-],[+]) + P([+],[+],[-])$$
$$= \left(\tfrac{1}{2}\right)^3 + \left(\tfrac{1}{2}\right)^3 + \left(\tfrac{1}{2}\right)^3 = 3 \times \left(\tfrac{1}{2}\right)^3 = \tfrac{3}{8},$$

and

$$P(3[+]) = P([+],[+],[+]) = \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} = \left(\tfrac{1}{2}\right)^3 = \tfrac{1}{8}.$$

Adding these probabilities together gives

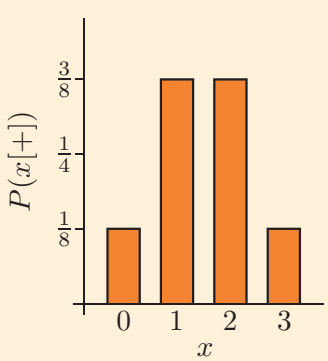
$$P(0[+]) + P(1[+]) + P(2[+]) + P(3[+]) = \tfrac{1}{8} + \tfrac{3}{8} + \tfrac{3}{8} + \tfrac{1}{8} = 1,$$

as required.

(b)

Sample of size 3

Number of [+]s, x	0	1	2	3
$P(x[+])$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$



Probability distribution for a random sample of size 3

Solution to Activity 25

(a) Now $n = 4$.

$$P(0[+]) = {}^4C_0 \times \left(\frac{1}{2}\right)^4 = 1 \times \left(\frac{1}{2}\right)^4 = \frac{1}{16}.$$

$$P(1[+]) = {}^4C_1 \times \left(\frac{1}{2}\right)^4 = \frac{4}{1} \times \left(\frac{1}{2}\right)^4 = \frac{4}{16} = \frac{1}{4}.$$

$$P(2[+]) = {}^4C_2 \times \left(\frac{1}{2}\right)^4 = \frac{4 \times 3}{2 \times 1} \times \left(\frac{1}{2}\right)^4 = \frac{6}{16} = \frac{3}{8}.$$

$$P(3[+]) = {}^4C_3 \times \left(\frac{1}{2}\right)^4 = \frac{4 \times 3 \times 2}{3 \times 2 \times 1} \times \left(\frac{1}{2}\right)^4 = \frac{4}{16} = \frac{1}{4}.$$

$$P(4[+]) = {}^4C_4 \times \left(\frac{1}{2}\right)^4 = \frac{4 \times 3 \times 2 \times 1}{4 \times 3 \times 2 \times 1} \times \left(\frac{1}{2}\right)^4 = \frac{1}{16}.$$

Probability distribution for a random sample of size 4

Number of [+]s, x	0	1	2	3	4
$P(x[+])$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$

The probabilities are positive and

$$\frac{1}{16} + \frac{1}{4} + \frac{3}{8} + \frac{1}{4} + \frac{1}{16} = \frac{1 + 4 + 6 + 4 + 1}{16} = 1.$$

(b) From the table, the probability that two of the selected values lie above the median (and hence two lie below it) is $\frac{3}{8}$.

If all selected values lie on the same side of the population median, they must be either all above or all below it. Applying the addition rule for mutually exclusive events, we obtain

$$\begin{aligned} P(4[+] \text{ or } 0[+]) &= P(4[+]) + P(0[+]) \\ &= \frac{1}{16} + \frac{1}{16} = \frac{1}{8} \quad (= 0.125). \end{aligned}$$

Solution to Activity 26

$$\begin{aligned} P(\text{number of [+]s equals 9}) &= {}^{12}C_9 \times \left(\tfrac{1}{2}\right)^{12} \\ &\simeq \frac{12 \times 11 \times 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4}{9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1} \times 0.000\,244 \\ &= \frac{12 \times 11 \times 10}{3 \times 2 \times 1} \times 0.000\,244 \simeq 0.054. \end{aligned}$$

Solution to Activity 27

There is no definitive answer to this question. Your answer may well differ from that of the module team, but it should follow the same pattern.

In the population of all schools, by definition, half will have a truancy rate above the median and half will have one below it. As a random sample is representative of the population, we might expect this property to apply in the sample to some extent.

(a) If all 12 values were above the assumed median, we should be very surprised. It is much more likely that the hypothesised median is not the real median. The module team would conclude that the hypothesis is almost certainly wrong.

(b) Again, this outcome seems very extreme, as only 1 value is above the hypothesised median. Again, the module team would conclude that the hypothesis is almost certainly wrong.

(c) This result is just what we would expect if the hypothesised median is equal to the true median. Thus the module team would conclude that the hypothesis is quite possibly true.

(d) With 7 values above and 5 below, we are only 1 value different from half above and half below. If we drew another sample, it could easily have 6 above and 6 below, or perhaps 5 above and 7 below. The module team would again conclude that the hypothesis is quite possibly true.

(e) Here you should be doubtful. The result is not very extreme, like cases (a) and (b), or very close to what we would expect, like cases (c) and (d). One could argue for (ii) that the hypothesis is probably wrong, though the module team would choose (i) that the hypothesis is quite possibly true. Some other phrase, such as 'the hypothesis is a little unlikely' would capture the view of the module team better.

Solution to Activity 28

(a) Since the outcomes are mutually exclusive we can use the addition rule to find the required probability. It is

$$\begin{aligned} &P(0[+]) + P(1[+]) + P(11[+]) + P(12[+]) \\ &\quad \simeq 0.000 + 0.003 + 0.003 + 0.000 \\ &\quad = 0.006. \end{aligned}$$

(b) The answer shows that the probability of obtaining one of these outcomes, when the population median is 0.98%, is extremely small. So if we observed one of these outcomes, it would seem sensible to reject the hypothesis and conclude that the population median is not equal to 0.98%.

Solution to Activity 29

(a)
$$P(0[+]) + P(1[+]) \simeq 0.000 + 0.000 = 0.000,$$

$$P(0[+]) + P(1[+]) + P(2[+]) \simeq 0.000 + 0.003 = 0.003,$$

$$P(0[+]) + P(1[+]) + P(2[+]) + P(3[+]) \simeq 0.003 + 0.014 = 0.017,$$

$$P(0[+]) + P(1[+]) + P(2[+]) + P(3[+]) + P(4[+]) \simeq 0.017 + 0.042 = 0.059.$$

(b) The probability of 0, 1, 2 or 3 values above the population median is the largest probability below 0.025.

(c) The 5% most extreme outcomes are 0[+], 1[+], 2[+], 3[+] and also 12[+], 13[+], 14[+] and 15[+]. The last four are the same as 0[−], 1[−], 2[−] and 3[−]. We should reject the hypothesis if there are three or fewer values on one side of the assumed median. The critical value at the 5% level for a sample of size 15 is 3.

(d) This corresponds to 1[+] and 14[−]. The smaller value is 1. As 1 is less than 3, we should reject the hypothesis at the 5% significance level.

(e) The smaller value is 5. As 5 is greater than 3, we should not reject the hypothesis at the 5% significance level.

(f) The smaller value is 3. As this is equal to the critical value, we should reject the hypothesis at the 5% significance level.

Solution to Activity 30

(a) From Table 8, the critical value for a sample of size 21 is 5. So the hypothesis would be rejected at the 5% significance level if the outcome were any of 0[+], 1[+], 2[+], 3[+], 4[+], 5[+], 0[−], 1[−], 2[−], 3[−], 4[−] or 5[−].

(b) The critical value for a sample of size 32 is 9. So the hypothesis would be rejected at the 5% significance level if the sample contained nine or fewer values above the assumed median, or nine or fewer values below the assumed median.

(c) The critical value for a sample of size 7 is 0. This means that the hypothesis would be rejected only if the sample contained 0[+] or 0[−]; in other words, it would be rejected only if all sample outcomes were on the same side of the assumed median.

Solution to Activity 31

(a) The median truancy rate for small schools in the East of England is 0.98%.

(b) Just 4 outcomes lie above the assumed median, and 19 outcomes lie below it. The test statistic is 4, the smaller of these numbers.

(c) From Table 8 (Subsection 4.1), the critical value for a sample of size 23 is 6.

(d) Since 4 is less than 6, we should reject the hypothesis at the 5% significance level and decide that the median truancy rate for small schools in the East of England is probably not equal to 0.98%. We can take our conclusion a little further. Since there are 19 values below 0.98% and only 4 values above it, we can conclude that the median truancy rate for small schools is probably *less than* 0.98%.

Solution to Activity 32

There are 28 outcomes (differences) and, if the median difference between the hybrids were 0 we would, on average, expect half the differences to be positive and half to be negative. In fact, only 7 of the differences are positive, while 21 of them are negative.

The smaller of these values is 7, so 7 is the value of the test statistic.

From Table 8 (Subsection 4.1), the critical value for a sample of size 28 is 8.

Hence the hypothesis is rejected at the 5% significance level. Hybrid B appears to give a higher yield than Hybrid A.

Solution to Activity 33

(a) The median truancy rate for schools in Essex is 0.98%.

(b) There are ten outcomes above the assumed median, and five outcomes below it. The test statistic is 5, the smaller of these numbers.

(c) From Figure 9, the p-value equals
$2 \times (0.000 + 0.000 + 0.003 + 0.014 + 0.042 + 0.092) = 2 \times 0.151 = 0.302$.

(d) The p-value is 0.302 (which is quite large), so there is little evidence against the hypothesis that the median truancy rate is 0.98% in schools in Essex.

Solution to Activity 34

(a) The smaller of 11 and 1 is 1, so we want

$$\begin{aligned} &P(0[-]) + P(1[-]) + P(0[+]) + P(1[+]) \\ &\quad = 2 \times [P(0[+]) + P(1[+])] \\ &\quad = 2 \times (0.000 + 0.003) = 0.006, \end{aligned}$$

from Figure 6.

The p-value is 0.006, which is very small, so there is strong evidence that the median truancy rate for academies is not 0.98%. It seems likely that their truancy rate is higher than that.

(b) Now the test statistic is 3.

$$\begin{aligned} &2 \times [P(0[+]) + P(1[+]) + P(2[+]) + P(3[+])] \\ &\quad = 2 \times (0.000 + 0.000 + 0.003 + 0.014) = 0.034, \end{aligned}$$

from Figure 9.

The p-value is 0.034, so there is moderate evidence that the median truancy rate for secondary academies is not 0.98%. The sample data suggest that their truancy rate is lower than that.

Solution to Activity 35

(a) Four values are above 1.0, three values are equal to 1.0, and 16 values are below 1.0.

(b) The test statistic is 4 (the smaller of 4 and 16). Originally there were 23 observations. We discard the three observations that tied with 1.0, leaving a sample size for the test of 20.

(c) From Table 8 (Subsection 4.1), the critical value at the 5% significance level for a sample size of 20 is 5. As $4 < 5$, we reject the hypothesis at the 5% significance level. The data provide moderate evidence that the median truancy rate of small schools in the East of England is not 1.0% – the median rate appears to be smaller than that.

Solution to Activity 36

(a) Nineteen values are above 19.7, three values are equal to 19.7, and eight values are below 19.7.

(b) The test statistic is 8 (the smaller of 19 and 8). Originally there were 30 observations. We discard the 3 observations that tied with 19.7, leaving a sample size for the test of 27.

(c) From Table 8 (Subsection 4.1), the critical value at the 5% significance level for a sample size of 27 is 7. As $8 > 7$, we do not reject the hypothesis at the 5% significance level. The data provide no clear evidence of climate change. (The p-value is actually 0.052, quite close to 5%, so there is some evidence that the stemplot data are not a random sample from the full set of data.)

Solutions to exercises

Solution to Exercise 1

This is a suggested ranking of the events, from the most likely to the least likely.

D. Death and taxes. (This gets top spot on the basis of the quote, 'the only things certain in life are death and taxes'.)

F. You get exactly three heads when you toss a coin five times. (This has a 10-in-32 chance of occurring – as you will be able to calculate using the results in Section 3.)

A. A husband and wife find they were born on the same day of the week. (This is a one-in-seven chance.)

H. It snows in London on Christmas day. (Data from 1950–2006 suggests this is about a 6-in-100 chance.)

E. A mother's first pregnancy results in twins. (Data for the United States suggest this is about a 3-in-100 chance.)

G. A chicken egg has two yolks. (The British Egg Information Service give this as less than 1-in-1000 chance.)

B. England will win the next football World Cup. (It's probably not as likely as getting an egg with two yolks.)

I. You will be struck by lightning next year. (If you live in the United States, the probability is about 1 in 280 000.)

C. A Brazilian team will win the next football European Cup. (As Brazil is not in Europe, a Brazilian team cannot play in the European Cup.)

As in Activity 6, your order may be slightly different from this, but it should be fairly similar.

Solution to Exercise 2

(a) The probability that the man has blue eyes is

$$\frac{\text{number of men with blue eyes}}{\text{total number of men}} = \frac{2811}{6800} \simeq 0.413.$$

(b) The probability that the man has brown hair is

$$\frac{\text{number of men with brown hair}}{\text{total number of men}} = \frac{2632}{6800} \simeq 0.387.$$

(c) The probability that the man has blue eyes and brown hair is

$$\frac{\text{number of men with blue eyes and brown hair}}{\text{total number of men}} = \frac{807}{6800} \simeq 0.119.$$

(d) The categories 'brown hair' and 'black hair' are mutually exclusive, as a man cannot have both. Hence the probability that the man has

brown hair or black hair is

$$\frac{\text{number of men with brown hair} + \text{number of men with black hair}}{\text{total number of men}}$$

$$= \frac{2632 + 1223}{6800} \simeq 0.567.$$

(e) The probability that the man does not have black hair is

1 − probability that he does have black hair

$$= 1 - \frac{\text{number of men with black hair}}{\text{total number of men}}$$

$$= 1 - \frac{1223}{6800} \simeq 1 - 0.180 = 0.820.$$

Solution to Exercise 3

If A is the event that Sue goes to the hockey match on Saturday and B is the event that her team wins, then A and B are statistically independent events (assuming that whether or not Sue does go to watch the match has no effect on how likely it is that her team wins), so

$$P(A \text{ and } B) = P(A) \times P(B) = 0.3 \times 0.4 = 0.12.$$

That is, the probability that Sue will watch her team play on Saturday and they will win is 0.12.

Solution to Exercise 4

(a) The number of ways of choosing three flags from seven flags in a specified order is

$$7 \times 6 \times 5 = 210.$$

Hence 210 different signals can be made.

(b) When a signal is flying there are four flags left in the box. The number of ways of choosing four flags from seven flags when order does not matter is

$${}^7C_4 = \frac{7 \times 6 \times 5 \times 4}{4 \times 3 \times 2 \times 1} = \frac{840}{24} = 35.$$

Hence there are 35 different combinations of flag that could be left in the box when a signal is flying.

Solution to Exercise 5

The probability of obtaining a head is $\frac{1}{2}$. Hence

P(three heads in five tosses of a coin)

$$= {}^5C_3 \times \left(\frac{1}{2}\right)^5 = \frac{5 \times 4 \times 3}{3 \times 2 \times 1} \times \frac{1}{32} = 10 \times \frac{1}{32} = 0.3125.$$

(This is the same as
P(three observations out of five are above the median).)

Solution to Exercise 6

The dealer's claim would be supported if the following were true.

The median petrol consumption of the car is 37 miles per gallon.

To determine the test statistic, we count the number of values above and below 37.0. There are 8 values above and 26 below. So the test statistic is 8.

From Table 8 (Subsection 4.1), the critical value for a sample of size 34 is 10. Since 8 is less than 10, we can reject the hypothesis at the 5% significance level and conclude that the median petrol consumption of the car is unlikely to be 37 miles per gallon. After looking at the sample, we can infer that the petrol consumption is probably less than 37 miles per gallon, and so the dealer's claim is not supported.

Solution to Exercise 7

There are 15 mice, and if the presence/absence of a mirror did not influence cage-preference, we would, on average, expect half the mice to spend more time in the cage with a mirror. In fact, only 3 of them did so, while 12 spent more time in the cage without a mirror.

The smaller of these values is 3, so 3 is the value of the test statistic.

From Table 8 (Subsection 4.1), the critical value for a sample of size 15 is 3.

Hence the hypothesis is rejected at the 5% significance level. We can conclude that mice appear to prefer a cage without a mirror.

Solution to Exercise 8

Three mice spent more time in the cage with the mirror, and 12 spent more time in the cage without a mirror.

The smaller of these values is 3, so 3 is the value of the test statistic.

From Figure 9, the p-value equals

$$2 \times (0.000 + 0.000 + 0.003 + 0.014) = 0.034.$$

This is moderately small – the p-value is between 0.01 and 0.05. Thus there is moderate evidence against the hypothesis. That is, there is moderate evidence that mice prefer a cage without a mirror.

Solution to Exercise 9

(a) Three values are above 0, four values are equal to 0, and nine values are below 0.

(b) The test statistic is 3 (the smaller of 3 and 9). Originally there were 16 observations. We discard the four observations that tied with 0, leaving a sample size for the test of 12.

(c) From Table 8 (Subsection 4.1), the critical value at the 5% significance level for a sample size of 12 is 2. As $3 > 2$, we do not reject the hypothesis at the 5% significance level. The data provide no clear evidence of methadone changing depression status.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Cover image: Minxlj/www.flickr.com/photos/minxlj/422472167/. This file is licensed under the Creative Commons Attribution-Non commercial-No Derivatives Licence http://creativecommons.org/licenses/by-nc-nd/3.0/

Introduction, cartoon (sharks and statistics): www.cartoonstock.com

Subsection 1.2 figure in Activity 3: Library of Congress

Subsection 1.2 cartoon (hula hoops): www.causeweb.org

Exercises on Section 1, cartoon (lightning): www.causeweb.org

Section 2 cartoon (fortune-teller): www.cartoonstock.com

Subsection 2.1 figure: ICMA Photos / This file is licensed under the Creative Commons Attribution-Share Alike Licence http://creativecommons.org/licenses/by-sa/3.0/

Subsection 2.3, figure by Activity 15: Taken from http://cdn.foodbeast.com

Subsection 3.1 photo (Programme Committee): The International Statistics Institute

Subsection 3.1 photo (combinations of clothes): Taken from http://makingthingsbeautifulagain.blogspot.co.uk/2012/11/project-333-season-1-list.html

Subsection 4.1, figure by Activity 32: Alternative Heat / Flickr.com

Exercises on Section 4, figure by Exercise 7: http://about.me/rafaeltronquini

Subsection 5.2, figure by Activity 36: George Jansoone / This file is licensed under the Creative Commons Attribution-Share Alike Licence http://creativecommons.org/licenses/by-sa/3.0/

Subsection 5.3 cartoon (data in perspective): www.causeweb.org

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked the publishers will be pleased to make the necessary arrangements at the first opportunity.

Unit 7

Factors affecting reading

Introduction

In Unit 6, we met some statistical techniques which enabled us to compare the truancy rate in large secondary schools in the East of England with the general truancy rate experienced by all secondary schools in the East of England. The method employed was to analyse sample data and use the results of the analysis to make inferences about the population from which the sample was drawn. In particular, we saw how the sign test enabled us to decide whether or not to reject, at the 5% significance level, the hypothesis that the population median takes a particular value.

Medians are just one measure of location. In this unit we return to hypotheses about location, and you will meet another hypothesis test, the 'z-test', which concerns *means*, rather than medians. For much of the unit we will be dealing with situations where we have a sample from a single population, as in Unit 6. We will then develop the ideas of hypothesis testing so as to compare *two populations* in terms of their locations. This involves setting up a hypothesis about the locations of the two populations (means, here, rather than medians) – the most common hypothesis is that the locations of the two populations are equal. A random sample of data is taken from each population, and these data are analysed to see whether or not to reject the hypothesis. Such tests are called 'two-sample tests', in contrast to 'one-sample tests' in the case of one population.

The emphasis will be on the development of statistical techniques, and, as in Unit 6, we shall explore many of the ideas in the context of a question taken from the general area of education. This time we shall be looking at the achievement of 7- and 8-year-old children in reading:

What factors affect a child's reading ability?

Section 1 starts with a brief discussion of this question. We shall then look at an available source of data and identify what aspects of the general question we can consider.

The next step will be to define specific questions of interest and use them to set up appropriate hypotheses. We will then begin to develop an appropriate sample statistic – a test statistic – with which to perform our hypothesis tests, by revisiting the idea of sampling distributions in Section 2. This notion will lead us to consider a particular distribution known as the 'normal distribution'. In Section 3, we look closely at this distribution, which is of great importance in statistics.

You were introduced to sampling distributions in Unit 4; they were used somewhat implicitly in Unit 6.

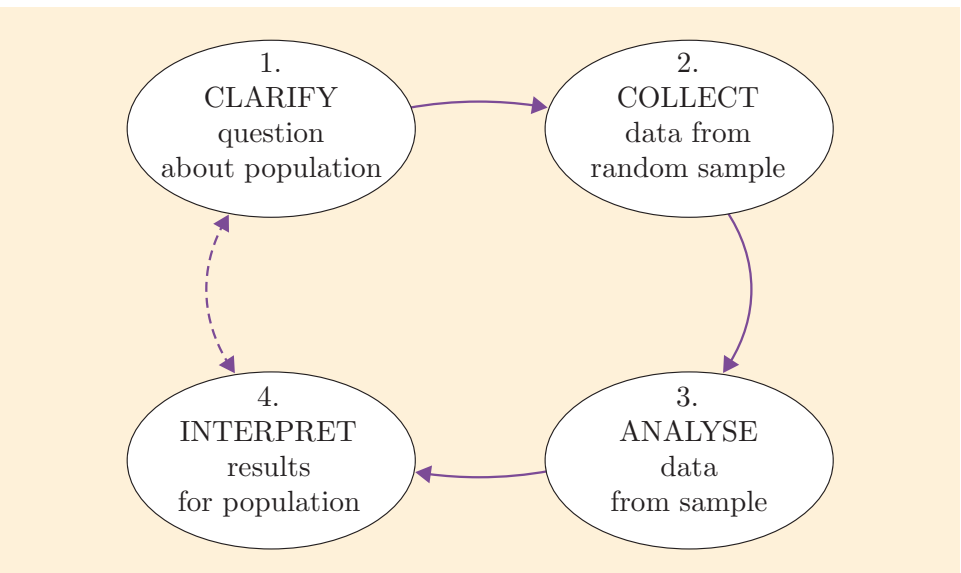
In Section 4, we go on to consider how the normal distribution helps us to define a usable test statistic, along with its sampling distribution. Section 5 is concerned with the application of the resulting z-test to the analysis of a sample of data from one population. Section 6 extends these ideas to investigate the difference between the means of two populations. One important aspect of these z-tests is that they are suitable only for dealing with (quite) large samples of data.

In Section 7, you will use Minitab to perform z-tests and learn to interpret the resulting p-values. Section 8 draws some conclusions about the educational question raised in the first section, and makes some general points about z-tests.

Section 7 directs you to the Computer Book. You are also guided to the Computer Book at the end of Subsections 3.1 and 3.3.

1 Clarifying the question

In Unit 6, the modified modelling diagram was introduced.



The modified modelling diagram (Figure 2 from Unit 6)

In this section, you are going to consider the first two stages of the modified modelling process: *clarify question* and *collect data*.

1.1 The question to be clarified

The question *What factors affect a child's reading ability?* is rather too general for us to attempt to answer it straight away. We need to make it more explicit. The first step is to understand what is meant by *reading ability*.

Activity 1 *Measuring reading ability*

How would *you* measure reading ability? Write down two or three measures of reading ability of 7- and 8-year-old children that you might use.

There are various different reading tests available to teachers, and they normally combine several of the measures mentioned in the solution to Activity 1. We shall be using data that have already been collected for us, so the measures used have already been defined.

A reading class

The next step is to consider the factors that might affect a child's reading ability.

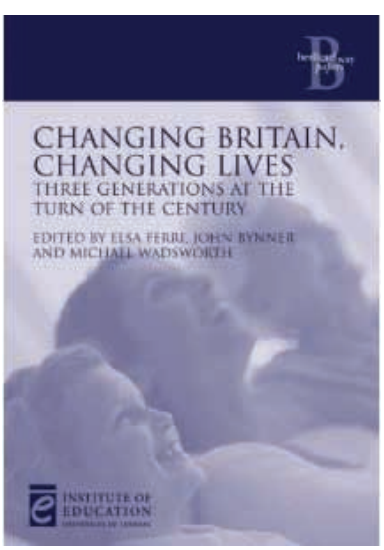
Activity 2 *Factors affecting a child's reading ability*

Write down some factors that you think might affect a child's ability to read.

The data that we shall use to explore this area will not allow exploration of all these factors. Therefore the data have to be examined before a decision can be made as to which factors can be explored and what questions can be asked.

1.2 The data to be used

We shall be looking at the population of British children aged 7 and 8. The sample we shall use consists of 7- and 8-year-old children of a certain group of parents defined as follows. At least one of the child's parents – a 'cohort member' – was born in a particular week in April 1970, resides in Great Britain, and has been part of a long-term study known as the British Cohort Study (BCS). There were more than 17 000 such people.



A book based on the results of the BCS and similar studies

The BCS had its origins in what was called the British Births Survey, which was originally designed to examine the social and biological characteristics of the cohort member's mother. That study looked at neonatal morbidity, and its results were compared with those of a similar, earlier study, the National Child Development Study, carried out in 1958. 'Neonatal morbidity' refers to disease of the child (the cohort member) in its first month of life.

Since 1970, the aims of the BCS have broadened considerably. There have been eight follow-up surveys, or 'sweeps', carried out in 1975, 1980, 1986, 1996, 1999–2000, 2004–2005, 2008–2009 and 2012. The follow-up surveys attempted to trace the original sample and, in the case of the first two follow-ups, to include immigrants born in the same week as the original sample. Each follow-up survey looked at different areas of the original group's development into adulthood. The one which included reading skills of the cohort member's children, and from which the data we shall be using

have been taken, was the one carried out by the Centre for Longitudinal Studies at the Institute of Education, University of London, in 2004–2005. At that time, the age of the people under study was 34 years.

We are therefore going to concentrate on data relating to the reading ability of children who were 7 and/or 8 years old in 2004–2005 and whose parents were part of the BCS. The 2004–2005 sweep of the BCS provided data on 745 children aged 7 or 8 in total. Of these, only 679 were tested for their reading ability. It is this sample of 679 that we shall concentrate on in this unit. As we progress through the analysis, you will find that we shall be using sample sizes smaller than this, because not all the additional information needed was provided in the answers to the questionnaire. However, the sample sizes concerned will remain pretty large.

Activity 3 *Is it a random sample?*

Write down some reasons why this sample of children can or cannot be considered a random sample of the population of 7- and 8-year-old children in Great Britain in 2004–2005.

We shall return to the issue of randomness of the sample in Section 8, but for most of the unit, despite our doubts, we shall assume that it *is* acceptable to treat the sample as if it were a random sample.

We next need to consider what data we have available. Table 1 shows data on the first few children in the sample. In the first column is the child's reading ability as scored using a standard reading test called the BAS II Word Reading Ability Score, where BAS stands for 'British Ability Scales'. This value will be referred to simply as the child's 'reading score' in this unit. The second column gives the child's age in months.

BAS II was, in fact, updated to BAS 3 in 2011.

The remaining columns of Table 1 are in coded form; that is, they use simple numerically coded values to represent attributes of the child in place of more complicated wordings, ranges of numbers or exact numerical values. For example, the third column shows the gender of the child, coded as 1 for a boy, 2 for a girl. The fourth column again relates to age; this time whether the child is aged 7 or 8 is recorded. The fifth column, headed 'Parental education', actually shows whether the cohort member's partner/spouse finished full-time education by the age of 16 or at some age over 16; it is used here as a measure of the level of education of the child's parents. The sixth column shows the occupation of the child's father. The codes for the values in columns three to six are given beneath the main body of the table.

Table 1 Part of the dataset on reading from BCS 2004–2005

Reading score	Age (months)	Gender	Coded age	Parental education	Father's occupation
106	91	1	1	1	–
123	95	2	1	1	1
123	86	2	1	–	1
110	92	1	1	1	1
92	90	2	1	2	–
129	93	2	1	1	–
118	97	1	2	–	2
115	107	2	2	1	2
117	93	2	1	2	–
134	89	1	1	1	1
25	85	1	1	–	2
110	93	2	1	1	1
172	94	1	1	1	1
138	90	2	1	–	2
56	105	1	2	–	1
136	100	2	2	1	1
115	90	2	1	1	–
160	94	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮

(This data is copyright and owned by the Economic and Social Data Service.)

'Gender', 1: boy; 2: girl. 'Coded age', 1: 7 years old; 2: 8 years old. 'Parental education', 1: finished aged 16 or less; 2: finished aged over 16. 'Father's occupation', 1: managerial, technical, professional and skilled non-manual occupations; 2: skilled manual, partly skilled and unskilled occupations.

You will notice that in some cases information is missing in the sample data. This is to be expected, because some people either cannot or do not wish to answer specific questions in the questionnaire. The missing data will just be ignored for now, but we will return to a brief consideration of its possible effects in Section 8.

A number of factors may have an effect on a child's reading ability. With our choice of data, the factors we can consider are child's age, child's gender, parental education and father's occupation.

Consider 'child's gender' first. What precise question can we ask? It must, as usual, be about the appropriate population (that of British children aged 7–8 in 2004–2005) and not merely the sample. We might ask, *Within this population, do boys and girls differ in their reading ability?* But we should be more precise. As in Unit 6, we shall be looking for a difference in *location*, in this case between reading scores of boys and girls.

In the next subsection we shall be more precise about the particular measure of location to use, but for now a reasonably precise question is:

> *For British children aged 7–8 in 2004–2005, did boys' and girls' ability scores differ in location?*

Similar questions can be asked about the other factors.

> *For British children aged 7–8 in 2004–2005, did reading scores differ in location according to the level of parental education?*

> *For British children aged 7–8 in 2004–2005, did reading scores differ in location according to their father's occupation?*

The questions can also be made more focused. For instance, consider the first question again. Perhaps there is a difference between boys and girls aged 7, but no such difference for 8-year-olds. Because of possibilities like this, it may be more appropriate to consider the two age groups separately, asking

> *For British children aged 7 in 2004–2005, did boys' and girls' reading scores differ in location?*

as well as

> *For British children aged 8 in 2004–2005, did boys' and girls' reading scores differ in location?*

The questions on education and occupation could also be split up according to age, in a similar way.

1.3 Setting up the hypotheses

We shall try to answer most of these very specific questions by means of hypothesis tests. Let us then remind ourselves what is involved by referring back to the procedure for the sign test discussed in Unit 6, but setting it up in a more formal way.

We began by making a statement about the population of interest that we wished to test. In particular, this was the hypothesis that the population median was equal to a specified value, M. The hypothesis that the population median is equal to M is known as the **null hypothesis**. This hypothesis is usually denoted by the symbol H_0. Thus the null hypothesis in the case of the sign test can be stated precisely in the form

> H_0: Population median $= M$.

We then looked at the data to see if there was any evidence that the population median did not, in fact, equal M. If there is evidence against the truth of the null hypothesis, H_0, then we *reject* this hypothesis and we conclude that there is evidence that the population median is not equal to M. That the population median is not equal to M is called the **alternative hypothesis**. An alternative hypothesis is usually denoted by the symbol H_1. Thus, if we reject the null hypothesis

> H_0: Population median $= M$,

then we are left with the alternative hypothesis

H_1: Population median $\neq M$,

'$\neq$' is the symbol for 'is not equal to'.

and we say we are rejecting the null hypothesis *in favour of* the alternative hypothesis.

In Unit 6, a trial in a law court was used as an analogy to hypothesis testing. In that context, the null hypothesis is that 'the defendant is not guilty', while the alternative hypothesis is that 'the defendant is guilty'. If the evidence against the null hypothesis is sufficiently great, then the jury should reject that hypothesis in favour of the alternative hypothesis, and conclude that the defendant is guilty.

Returning to the questions concerning children's reading ability, the first step therefore is to set up the appropriate null and alternative hypotheses. As you might expect, these correspond to whether or not there is a difference in location in reading ability between two groups of children. We shall start with one of the questions on gender.

For British children aged 7 in 2004–2005, did boys' and girls' reading scores differ in location?

However, before defining the hypotheses, it is worth thinking again about the data. We actually have data on reading scores and gender for 396 children who are aged 7 (of these, 206 are boys and 190 are girls). Printing all 396 scores here would clearly be cumbersome and waste space. We can summarise the data as shown in Table 2.

Table 2 Summary statistics for data on reading scores of 7-year-old children

	Sample size	Sample mean	Sample standard deviation
Boys	206	109.31	27.671
Girls	190	113.42	25.464

(This data is copyright and owned by the Economic and Social Data Service.)

You may well be wondering why the summary measures are the mean, $\overline{x}$, and standard deviation, s, and not some other measures of location and spread, such as the median, M, and interquartile range, IQR. A minor reason is that the mean and standard deviation are commonly used in practice, so more people are familiar with them than with other measures. The main reason, though, is that $\overline{x}$ and s can be used to construct a reasonably simple test, in a way that M and IQR cannot.

Activity 4 *Calculating a mean and standard deviation*

The calculation of the mean was discussed in Subsection 1.3 of Unit 2 and the calculation of the standard deviation was discussed in Subsection 3.1 of Unit 3.

Because it is some time now since you worked with the sample mean ($\overline{x}$) and the sample standard deviation (s), here is a reminder of how to calculate these summary measures:

$$\overline{x} = \frac{\sum x}{n},$$

where $\sum x$ is the sum of all the sample values and n is the sample size; and

$$s = \sqrt{\text{variance}},$$

where the variance is $\dfrac{\sum (x - \overline{x})^2}{n - 1}$.

(a) Data on the first eighteen 7- and 8-year-olds taken from the BCS 2004–2005 results were given in Table 1 (in Subsection 1.2). Extract from that table the values of the reading scores for all the 7-year-old boys. What is the value of n for this small sample?

(b) Calculate $\overline{x}$ and s for the reading scores for 7-year-old boys that you extracted in part (a).

Having paused briefly to examine the sample data, we now move on. We still need to state the null and alternative hypotheses associated with the question

> *For British children aged 7 in 2004–2005, did boys' and girls' reading scores differ in location?*

in their precise forms. The null hypothesis will be

> H_0: For British children aged 7 in 2004–2005, the mean reading score for girls was equal to the mean reading score for boys.

As you will have noticed, H_0 is phrased in terms of the population means and not, for example, the population medians. The alternative hypothesis is naturally taken to be

> H_1: For British children aged 7 in 2004–2005, the mean reading score for girls was not equal to the mean reading score for boys.

The null and alternative hypotheses for other questions listed at the end of the previous subsection are similar.

The British Ability Scales reading score system gives overall mean test scores for different age groups in Great Britain. These overall mean test scores are given for quite finely defined age groups, from which the authors of this unit have come up with the following means for 7- and 8-year-olds: the population mean for 7-year-old children is 96, and for 8-year-olds it is 116. (Actually, these means come from very large samples of children and not the whole population, but in practice we can treat them as population means.) So a further appropriate question to ask about the data on reading scores for 7-year-old children, for example, is whether they are

consistent with a population mean of 96. In other words, we could test the following hypotheses:

H_0: For British children aged 7 in 2004–2005, the mean reading score was equal to 96

H_1: For British children aged 7 in 2004–2005, the mean reading score was not equal to 96.

In testing hypotheses about population medians in Unit 6, the next step was to define a quantity that we could calculate from the data that would help us to evaluate the truth or otherwise of the null hypothesis. In the law-trial analogy, this is the evidence. In the sign test, this quantity was the smaller of the numbers of [+]s and [−]s that the sample contains. (See Section 4 of Unit 6.) In general, in hypothesis testing, this quantity is called the **test statistic**. So now we need to find suitable test statistics to assess the hypotheses about children's reading abilities. Since these hypotheses are about population means or differences between population means, the obvious test statistics would involve sample means or the differences between sample means. But, as with the sign test in Unit 6, the awkward part involves finding what is called the sampling distribution of the test statistic; so in the next section we look again at sampling distributions.

Exercises on Section 1

Exercise 1 *Mean and standard deviation for 8-year-olds*

(a) Extract from Table 1 the values of the reading scores for all the 8-year-old children in the table. What is the value of n for this small sample?

(b) Calculate $\overline{x}$ and s for the reading scores for 8-year-old children that you obtained in part (a).

Exercise 2 *Parental education and occupation*

(a) Extract the values of the reading scores for all the children in Table 1 whose parent's age on finishing full-time education was 16 or less and whose father's occupation is managerial, technical, professional or skilled non-manual. What is the value of n for this small sample?

(b) Calculate $\overline{x}$ and s for the reading scores for the sample of children that you obtained in part (a).

Exercise 3 *Null and alternative hypotheses?*

Suggest null and alternative hypotheses for comparing the reading abilities of the 8-year-old children according to their gender.

2 Sampling distributions revisited

In Subsection 3.3 of Unit 4, you saw what is meant by the sampling distribution of the median of a sample, and what happened to such sampling distributions as the sample size increased. We now review these ideas, but rather than just repeating exactly what was done before, we look at the **sampling distribution of the mean** as opposed to that of the median.

As in Unit 4, in order to look at these sampling distributions precisely, we really need to know all the relevant information about the whole population. Nobody has information about the reading ability of all 7- and 8-year-old children in Great Britain, so we cannot work with data exactly like those from the BCS. Instead let us look at a population where we *do* have data on everyone, and investigate sampling distributions using that. The population is that of all students taking the examination for the Open University module *Exploring mathematics* (MS221) in a particular presentation. There were 1234 students in the presentation chosen, and their marks in the examination are displayed in Figure 1. This plot is very like a histogram with lines instead of bars. The numbers of students achieving each mark from 0 to 100 are given by the heights of the lines drawn at each mark. These heights are the same as the areas of the bars that would have been used on the histogram. But, in addition, the top ends of the lines have been joined together.

Histograms were introduced in Subsection 1.5 of the Computer Book.

This representation gives a good picture of the shape of the *population distribution* of examination marks of students on MS221 in one presentation.

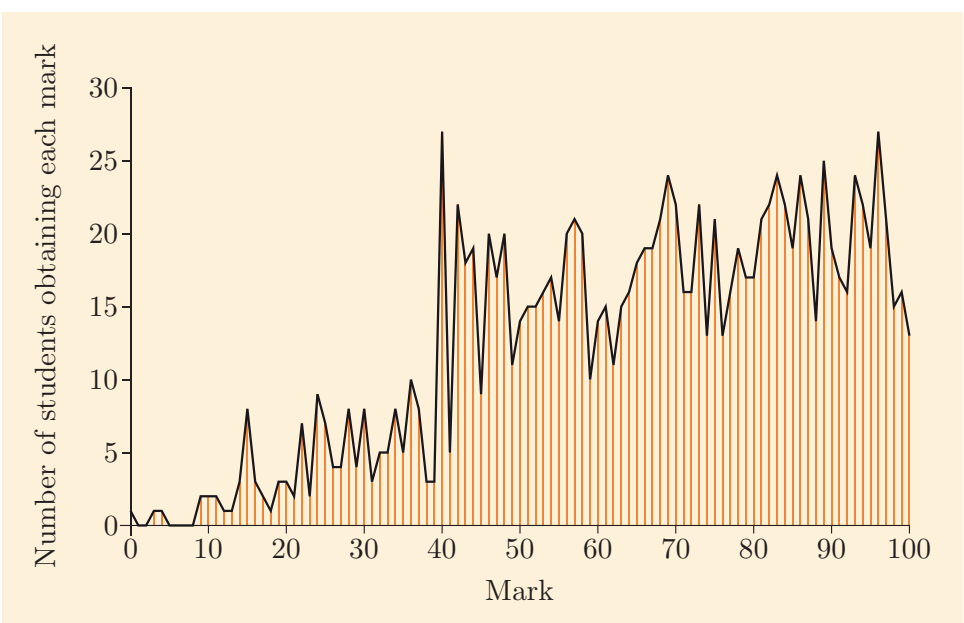


Figure 1 Numbers of students obtaining each examination mark in MS221

Now, there is a modification that we need to make. What will be important later are the *proportions* of students in the population gaining each mark. Thus instead of using the vertical axis to measure the actual number of students who obtained each mark in the examination, we want the population distribution to be described in terms of the proportion of students in the population who obtained each mark. We can do this simply by dividing each of the actual numbers represented in Figure 1 by the total number of students in the population (1234). Hence,

$$1 \text{ becomes } \frac{1}{1234} \simeq 0.0008,$$
$$2 \text{ becomes } \frac{2}{1234} \simeq 0.0016,$$
$$3 \text{ becomes } \frac{3}{1234} \simeq 0.0024,$$

and so on.

Activity 5 *From number to proportion*

The actual number of students scoring 75 marks in Figure 1 is 21. What proportion of students on MS221 in the presentation in question achieved 75 marks?

The result of changing from actual numbers to proportions is shown in Figure 2. Notice that Figure 2 looks just the same as Figure 1; only the scale on the vertical axis has changed.

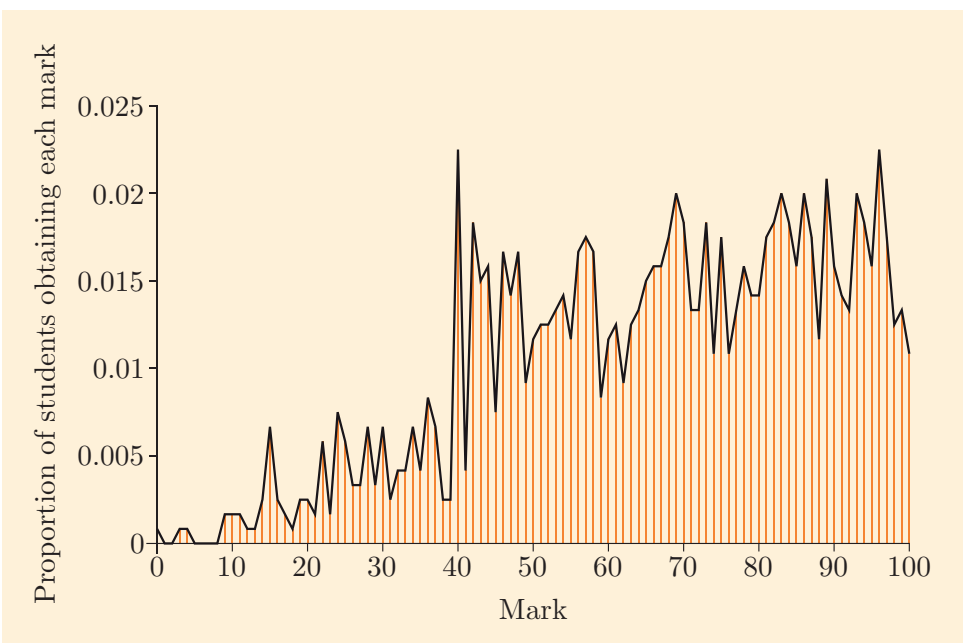


Figure 2 Proportions of students obtaining each examination mark in MS221

However, it is not the characteristics of the population distribution of exam marks, above, that we are interested in as such. Our focus is going to be on the sampling distributions of *means of random samples of exam marks taken from this population*. This is because we will be interested in testing hypotheses about the mean examination mark, such as

H_0: For students on MS221, the mean examination score is equal to 65

H_1: For students on MS221, the mean examination score is not equal to 65

or (using data from other years)

H_0: For students on MS221, the mean examination score for the current presentation is equal to the mean examination score for the previous presentation

H_1: For students on MS221, the mean examination score for the current presentation is not equal to the mean examination score score for the previous presentation.

Now we begin our investigation of the sampling distribution of the mean. Consider first all possible random samples of size 2 that we might select from the population data of 1234 examination marks. There is a great number of possibilities (760 761, to be precise!), and we cannot concisely picture *all* the sample values in every one of these possible samples. However, as in Unit 4, we *can* summarise each sample using a *summary measure*, and then picture these in the form of the sampling distribution of that summary measure. This time, as suggested above, we use the *sample mean* as our summary measure.

Activity 6 *Sample means of samples of size 2*

(a) Find the sample means of each of the following samples of size 2:

(i) 15, 35 (ii) 65, 77 (iii) 65, 52 (iv) 37, 80.

(b) The exam marks in the population, and hence in any sample, are all integers (whole numbers). Are the sample means of samples of size 2 necessarily integers? If not, what other kinds of value can these sample means take?

In this way it would be possible to calculate the sample mean for every one of the 760 761 possible samples of size 2. Different samples can give the same sample mean, as (iii) and (iv) in part (a) of Activity 6 illustrate. The sampling distribution records the *proportions* of all these samples with each value of the sample mean. A picture of this is shown in Figure 3. This represents the *sampling distribution of the mean for samples of size 2* from the population of exam marks. Here all the possible values of the sample mean $\overline{x}$ are indicated on the horizontal axis, and the vertical lines represent the heights of the bars that would be used for a histogram of the proportion of samples (out of 760 761 possibilities) which have each of these values as the sample mean. Notice that there are many more lines in

this diagram than there are in Figure 2. That is because this sample mean can take about twice as many values, integers and half-integers, as you saw in Activity 6, so the histogram can have twice as many bars. Joining the tops of the lines again provides us with a good picture of the *shape* of the distribution.

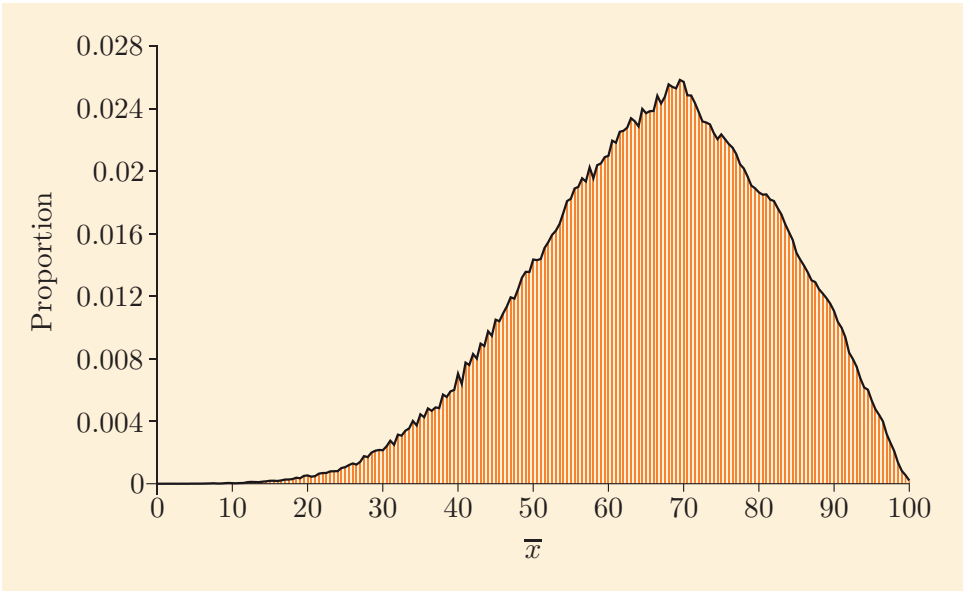


Figure 3 Sampling distribution of the mean for samples of size 2 from the population of MS221 exam marks

Activity 7 *Distribution of sample means of size 2*

What are the main features of the distribution of sample means of size 2 shown in Figure 3?

Let us now find out, as in Unit 4, what happens to the sampling distribution as the sample size increases. Let's first look at the sampling distribution of the mean for samples of size 3.

Activity 8 *Sample means of samples of size 3*

Find the sample mean of each of the following samples of size 3:

(a) 10, 20, 45 (b) 82, 24, 33 (c) 52, 61, 73 (d) 78, 64, 46.

Activity 8 indicates that there are even more possible values of the sample mean for samples of size 3 than there are for samples of size 2. This means that the vertical lines in the sampling distribution will be even closer together. For this reason, we stop plotting the lines and just concentrate on the *shape* of the distribution as indicated by the *tops* of the lines; we obtain the picture of the sampling distribution shown in Figure 4. In fact,

the 'joining' line shown in Figure 4 is made up of lots of very short lines, each one joining two adjacent vertical lines.

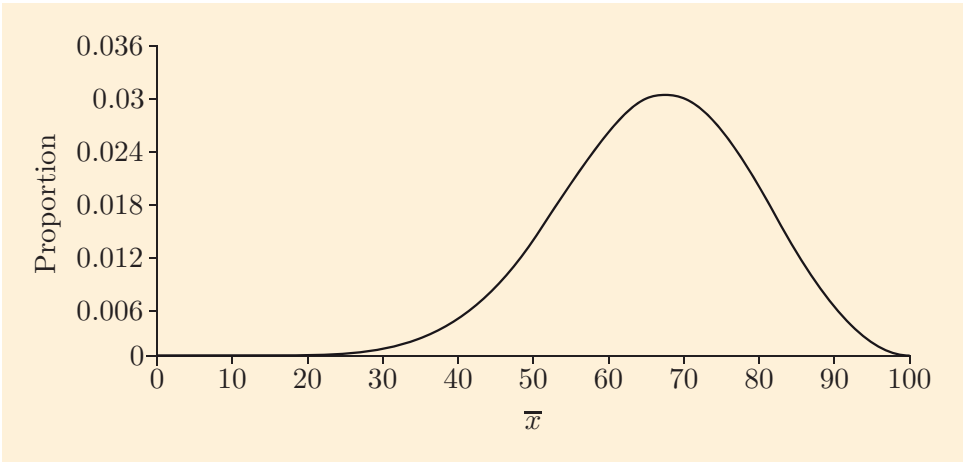


Figure 4 Sampling distribution of the mean for samples of size 3 from the population of MS221 exam marks

Activity 9 *Distributions of sample means of sizes 2 and 3*

How does the distribution of sample means of size 3 shown in Figure 4 compare with the distribution of sample means of size 2 shown in Figure 3?

Activity 10 *Distributions of sample means of sizes 3 and 5*

Figure 5 shows the distribution of sample means of size 5 from the population of MS221 examination marks. How does the distribution shown in Figure 5 compare with the distribution of sample means of size 3 shown in Figure 4?

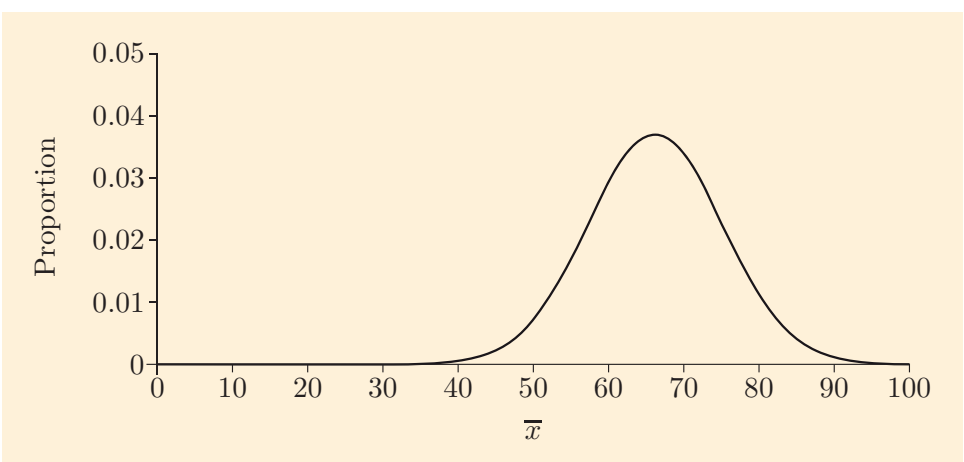


Figure 5 Sampling distribution of the mean for samples of size 5

Activity 11 *Distributions of sample means of larger sample sizes*

Figure 6 contains pictures of the sampling distributions of the mean for larger sample sizes. Notice that we have not indicated the scale on the vertical axes in Figure 6, but it is the same in each case. Describe the changes in shape of these distributions, as the sample size n increases.

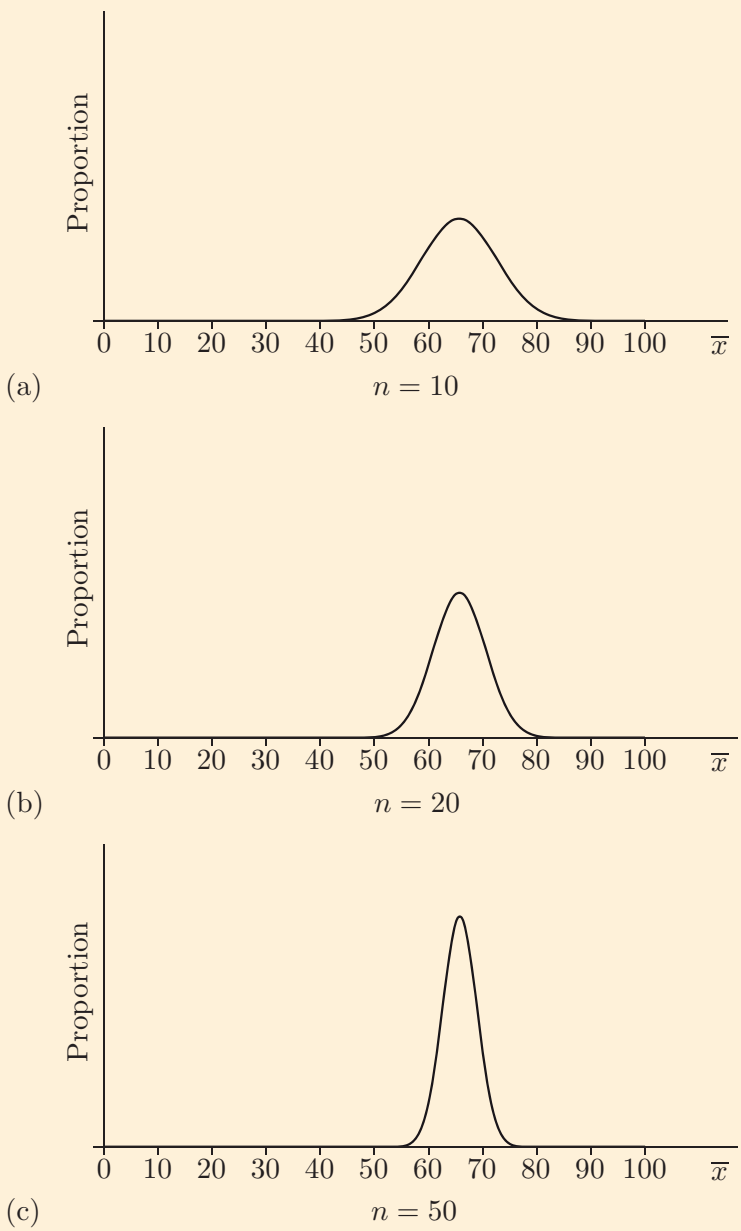


Figure 6 Sampling distributions of the mean for samples of size n

The common shape of the distributions in Figures 4, 5 and 6 is sometimes called a 'bell shape'. You can use the following picture of Big Ben to decide whether or not you agree that these distributions are bell-shaped!

Big Ben: bell-shaped?

Now the interesting thing about the sampling distribution of the mean is that it will nearly always be approximately bell-shaped (looking something like the above figures), *no matter what population distribution is taken as the starting point.* (The sampling distributions of some other quantities, such as the sample median, show similar features.)

Example 1 *Sampling distributions of means based on earnings data*

Figure 7 provides a rough picture of the population distribution of earnings of all full-time employees in the UK in 2011.

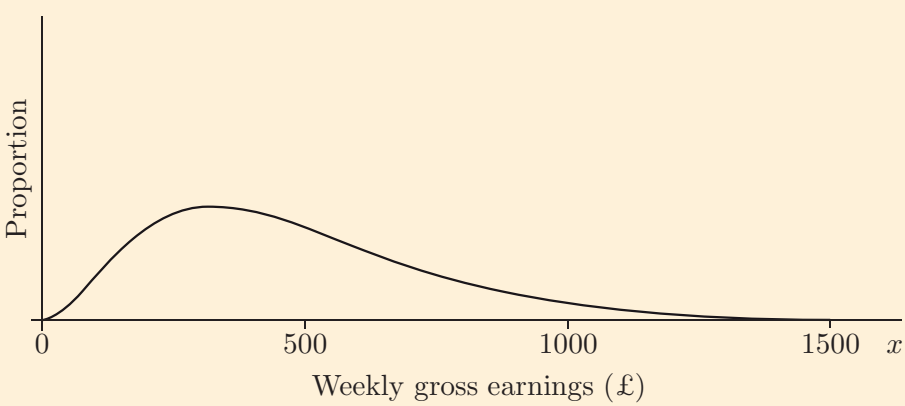


Figure 7 Population distribution of earnings of full-time employees

This population distribution is very smooth. The smoothness results from the fact that the population is extremely large and there are so many possible earnings that we can record. This means that the vertical lines representing the various adjacent proportions would be so close together that we could not distinguish between them and so, effectively, the line joining the tops is a smooth curve. The distribution is, however, clearly right-skew since it has a long tail to the right. This reflects the fact that while most employees earn a moderate to 'medium' wage, some employees earn considerably more, and a few earn very considerably more again.

Figure 8 contains pictures of the sampling distributions of the *mean* for samples of various sizes from this population distribution.

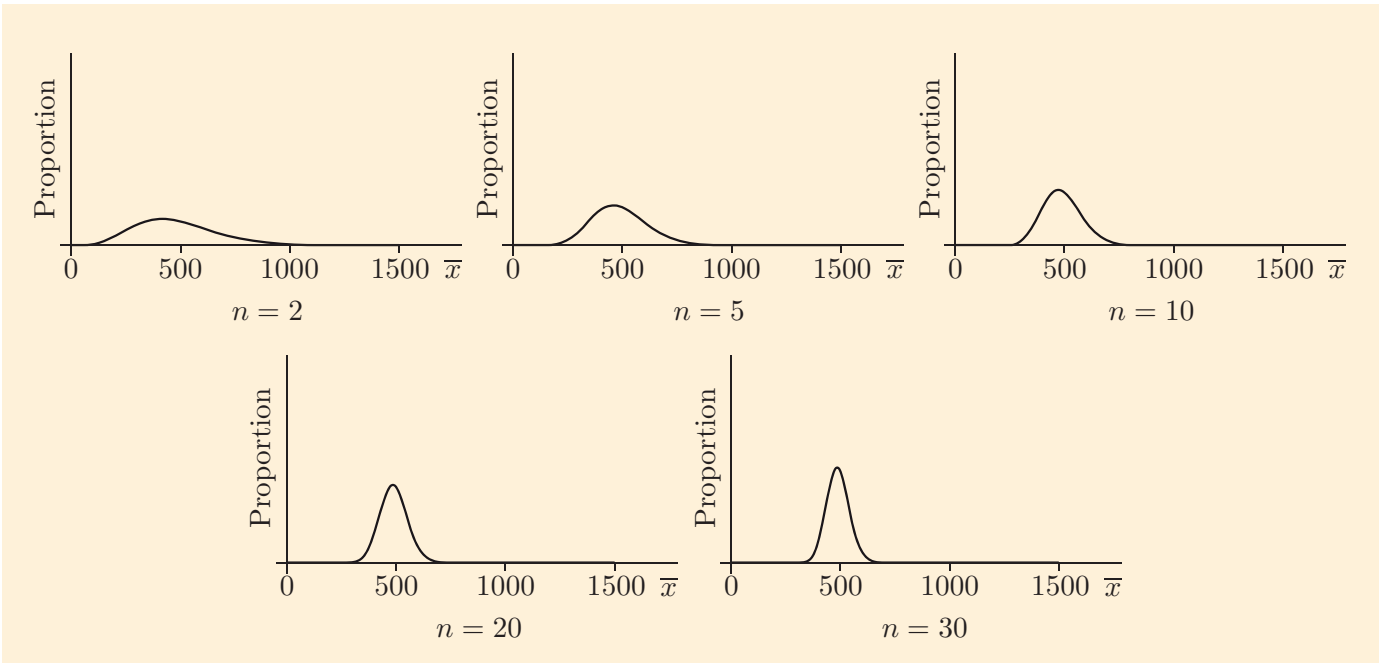


Figure 8 Sampling distributions of means based on earnings data

Activity 12 *Distributions of sample means of earnings data*

Describe the main changes in shape of the sampling distributions in Figure 8, as the sample size n increases.

So, again, we see from Example 1 and Activity 12 that even though the population distribution is skew, as the sample size n increases, the sampling distribution of the mean becomes more and more symmetric and bell-shaped.

What is surprising, though, is that if the sample size is large enough, the sampling distribution of the mean will nearly *always* be this sort of shape, no matter what shape the population distribution is.

The shape of sampling distributions of the mean

For most practical purposes, whatever the shape of the original population distribution, the sampling distributions of the mean for large enough sample sizes are always symmetric and bell-shaped.

These symmetric bell-shaped distributions that we obtain as sampling distributions for large enough values of n are called **normal distributions**.

As you will see in Section 3, these distributions have some very interesting properties which help us to develop the test statistic that we are working towards.

What is a large enough sample?

As a rough guide, you can assume that, whatever the population distribution, for sample sizes greater than 25, the sampling distribution of the mean will always be approximately normal, and in practice, we generally assume that it *is* normal.

In fact, the sampling distribution of the mean will actually be approximately normal for sample sizes (much) smaller than $n = 25$ for many population distributions. On the other hand, there are atypical population distributions for which the sampling distribution of the mean is not (approximately) normal. You will not deal with samples from such populations in M140. This allows us to rephrase a previously highlighted statement.

The shape of sampling distributions of the mean, rephrased

For most practical purposes, whatever the shape of the original population distribution, the sampling distributions of the mean for large enough sample sizes are always *approximately normal*.

Exercises on Section 2

Exercise 4 *Means of samples of size 2 from two small populations*

Consider the following two small populations of values:

Population A: 10 20 30 40 and Population B: 10 38 39 40

(a) Find the sample mean of each of the six different samples of size 2 that you can obtain from Population A. Make a very rough plot of the positions of the six sample means along the horizontal axis.

(b) Repeat what you did in part (a) for Population B.

(c) Compare the graphs you obtained in parts (a) and (b). Which of the two displays a more bell-shaped distribution of sample means? Can you think of a reason why this should be so?

Exercise 5 *Change in shape as sample size changes?*

The BCS sample with which we are concerned in this unit comprises a total of 679 reading scores (of 7- and 8-year-old children in 2004–2005). We will now *pretend* that this large sample of reading score values is actually the entire population of reading score values. Figure 9 contains pictures of the sampling distributions of the *mean* for samples of various sizes from the (pseudo-)population distribution of reading scores. Describe the changes in shape of these sampling distributions, as the sample size n increases.

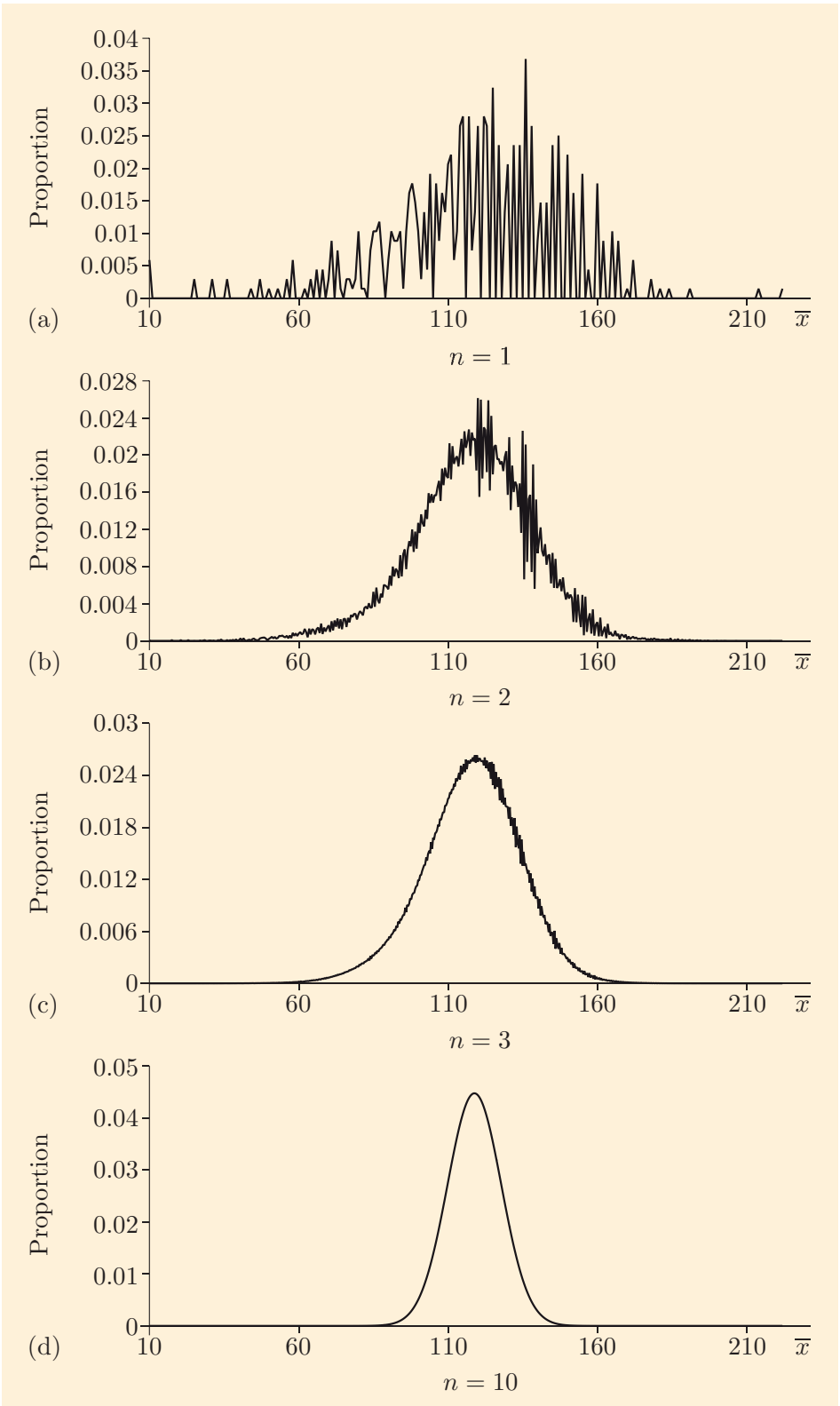


Figure 9 Sampling distributions of the mean as sample size changes

((a) This data is copyright and owned by the Economic and Social Data Service.)

3 Normal distributions

In Section 2 we saw that the sampling distribution of the mean is nearly always approximately normal, provided the sample size is sufficiently large. In this section we examine some of the properties of normal distributions and begin to discover just how important sampling distributions really are.

But first, we need to introduce some important new terminology. You are already familiar with the idea of a sample mean, $\overline{x}$:

$$\overline{x} = \frac{\sum x}{n} = \frac{\text{sum of sample values}}{\text{sample size}}.$$

In this section we shall also need to refer to the **population mean**. For a population of finite – but very large – size, N, this is calculated in exactly the same way, but using *all* the data values in the population. By convention it is labelled μ, so that

The symbol μ is the lower-case Greek letter 'mu', pronounced to rhyme with 'new'.

$$\mu = \frac{\sum x}{N} = \frac{\text{sum of population values}}{\text{population size}}.$$

Because N is often very large indeed, the population is often actually assumed to be of infinite size. For an infinite population, the population mean value, μ, is the mean of a truly enormous sample – the sample size must approach infinity.

There is a similar distinction between the sample standard deviation, s, and the **population standard deviation**, which is denoted by the symbol σ. (This is the lower-case version of the Greek letter 'sigma', which in upper-case form is $\sum$, but there is no connection between the ways that these two symbols are used here.) The formulas are

$$s = \sqrt{\frac{\sum(x-\overline{x})^2}{n-1}} = \sqrt{\frac{\sum x^2 - (\sum x)^2/n}{n-1}},$$

where the summations are over *sample* values, and

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}} = \sqrt{\frac{\sum x^2 - (\sum x)^2/N}{N}},$$

where the summations are over *population* values.

An important property to note is that σ is always a positive number.

Off duty from their work in statistics class μ and σ take a much needed Spring Break in their native Greece

3.1 Normal distributions: location and spread

Normal distributions are important in statistics for two different reasons. You met the first of these in Section 2: many sampling distributions of summary statistics are approximately normal for large enough samples. The other reason is that the distributions of many *populations* are approximately normal. One example is the population of men's heights that you will look at below. In Unit 10 you will see further examples of population distributions that are approximately normal.

Importance of the normal distribution

Normal distributions are important both as (approximate) population distributions, in some cases, and as (approximate) sampling distributions, in many more cases.

The distribution is called the *normal* distribution because it arises so commonly. Normal distributions are also called **Gaussian distributions** after the great German mathematician and scientist C.F. Gauss, who was instrumental in their development. They also appear in popular literature as the 'bell curve'.

Carl Friedrich Gauss
(1777–1855)

Carl Friedrich Gauss

Gauss (1777–1855) was a phenomenal mathematician – one of the most productive mathematicians ever. He made exceptional contributions in many fields, perhaps number theory most notably, but also astronomy, geometry, algebra, geophysics and, amongst others, statistics. During the early part of his career he took up the challenge of predicting where Ceres would be found. Ceres was a dwarf planet that had been observed in 1801 but which then disappeared behind the Sun and could not be found when it first reappeared. Gauss developed new methods of estimation and approximation to locate its position. He later published a monograph on the theory of the motion of small planets disturbed by large planets, and in this he introduced several important statistical concepts, including the normal distribution. It is for this reason that the normal distribution is also called the Gaussian distribution, though Gauss did not contribute most to the development of its properties. (The contribution of Laplace (1749–1827), for example, is greater.)

In this unit, we want to explore certain characteristics of normal distributions in order to apply them to sampling distributions. It would be possible to do this exploration using the sort of sampling distributions we met in the last section. However, the descriptions of what is going on tend to look rather complicated, because they involve means of sampling distributions of means. To make things clearer, the exploration is therefore done in the context of a normally distributed population.

Each normal distribution is a precise distribution defined by a mathematical formula involving the mean and standard deviation. We shall not need to use this formula in this module. But despite this mathematical precision, in practice the word ‘approximate’ is very important above. Real-world populations never have *exact* normal distributions in terms of the mathematical formula; but many are close enough to a normal distribution so that it makes sense to *treat* them as having normal distributions, in which case we say they are approximately normally distributed.

Figure 10 provides a picture of the population distribution of the heights of all men in Scotland in 2008, based on information given by the *Scottish Health Survey,* 2008.

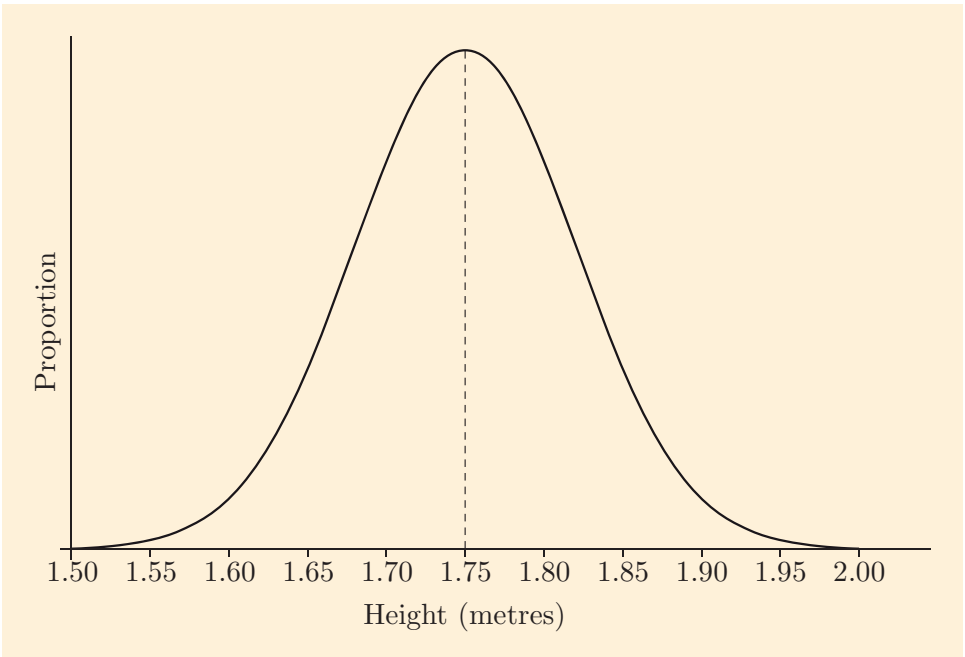


Figure 10 Population distribution of Scottish men's heights (in metres)

This population distribution is very smooth, symmetric and bell-shaped. For the rest of this section, we shall assume that the distribution is indeed normal.

The symmetry of the distribution means that the population mean height is the value corresponding to the mode (peak) of the distribution: about 1.75 metres. (In fact, as well as being the mode and the mean, this value is also the population median!) This characteristic applies more generally so that any normal distribution is symmetric about its mean μ.

Figure 11 shows normal distributions for different values of the mean μ and Figure 12 shows normal distributions for different values of the standard deviation σ.

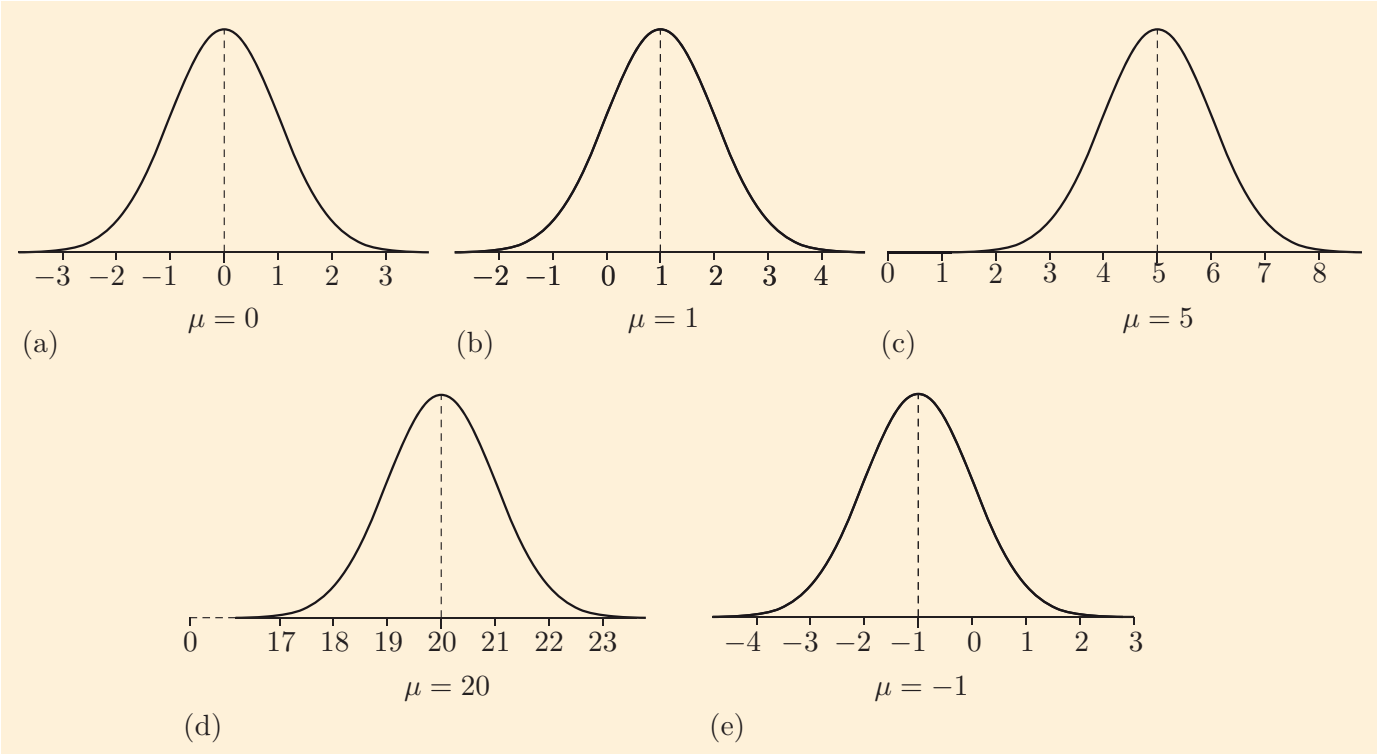


Figure 11 Normal distributions with different locations

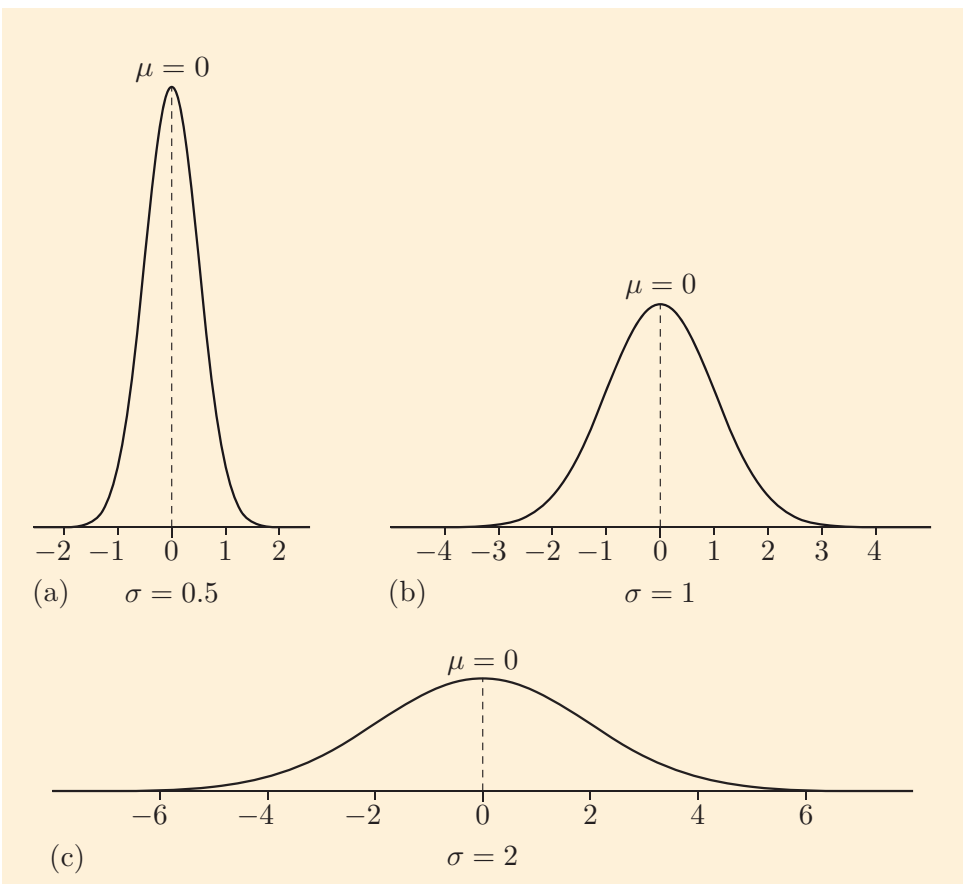


Figure 12 Normal distributions with different spreads

The location of a normal distribution on the horizontal axis depends on the value of its mean μ, as demonstrated by Figure 11.

As with any distribution, the spread of a normal distribution can be measured by the standard deviation of the population, σ. Thus a small value of σ means that the distribution is tightly clustered about the mean; the larger the value of σ, the more spread out the distribution will be – as demonstrated by Figure 12.

Activity 13 *What are μ and σ for this normal distribution?*

Figure 13 shows another normal distribution. By comparing it with Figures 11 and 12, can you identify the values of μ and σ for this normal distribution?

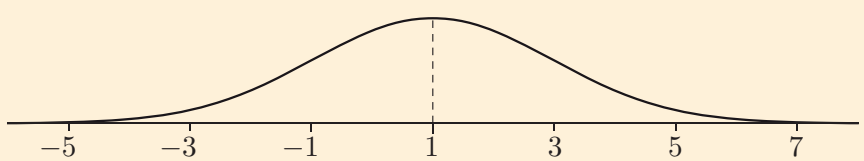


Figure 13 A normal distribution related to those in Figures 11 and 12

Location and spread of the normal distribution: 1

The normal distribution has location specified by the population mean, μ, and spread specified by the population standard deviation, σ.

You have now covered the material needed for Subsection 7.1 of the Computer Book.

You have also now covered the material related to Screencast 1 for Unit 7 (see the M140 website).

3.2 Normal distributions: relating means, standard deviations and plots

For a normal distribution, almost the whole of the distribution (about 99.7%) is contained within plus or minus three standard deviations of the mean. For example, the population distribution of Scottish men's heights (in metres) is normal with mean $\mu \simeq 1.75$ and standard deviation $\sigma \simeq 0.07$. Thus $3\sigma \simeq 0.21$, and so almost the whole of the distribution is contained within plus or minus 0.21 metres of the mean 1.75 metres (i.e. between $1.75 - 0.21 = 1.54$ metres and $1.75 + 0.21 = 1.96$ metres). You can check for yourself in Figure 14 – which is an annotated copy of Figure 10 – that this is indeed the case.

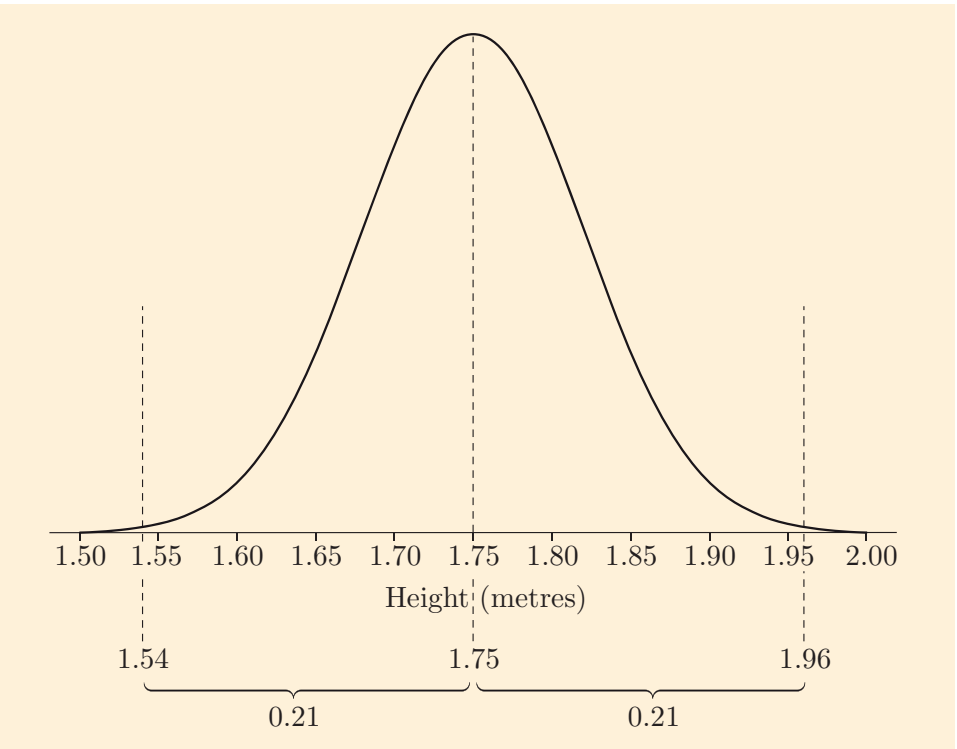


Figure 14 Annotated population distribution of Scottish men's heights

(Similar percentages are known for all other numbers of standard deviations; for instance, 95.4% of the distribution is contained within plus or minus two standard deviations, and 68.3% within plus or minus one standard deviation.)

Location and spread of the normal distribution: 2

The normal distribution has its mode at μ, and almost the whole of the normal distribution is contained between $\mu - 3\sigma$ and $\mu + 3\sigma$.

The links between the graph of a normal distribution and its mean and standard deviation suggest that a picture of the distribution can be used to obtain approximate values for its mean and standard deviation.

Example 2 *Approximate values for the mean and standard deviation*

The population distribution of a certain variable x is known to be normal. This distribution is pictured in Figure 15.

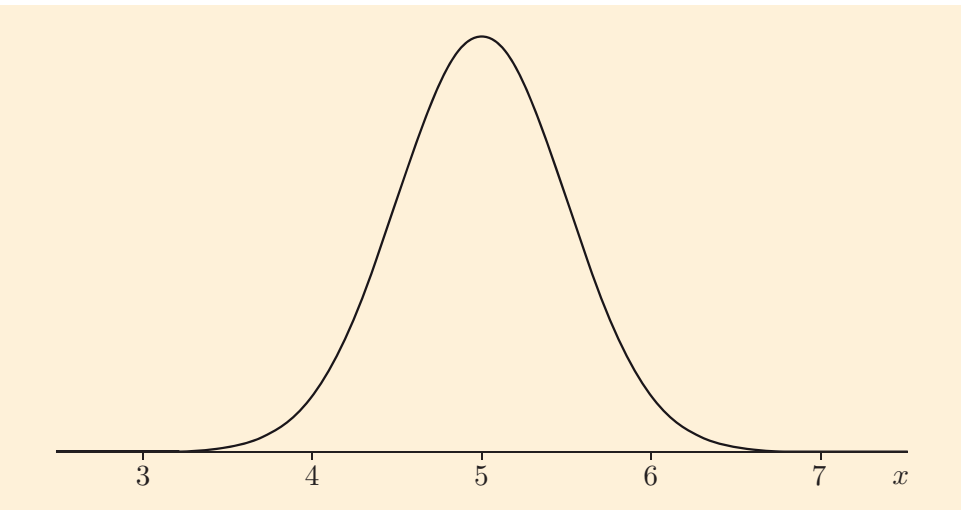


Figure 15 A normal distribution

The mode of this normal distribution occurs at about $x = 5$. This means that the population mean must be approximately equal to 5. So, $\mu \simeq 5$. We say *approximately* equal because μ may not be exactly equal to 5. It could be 5.1 or 4.9; it is impossible to give an exact value here.

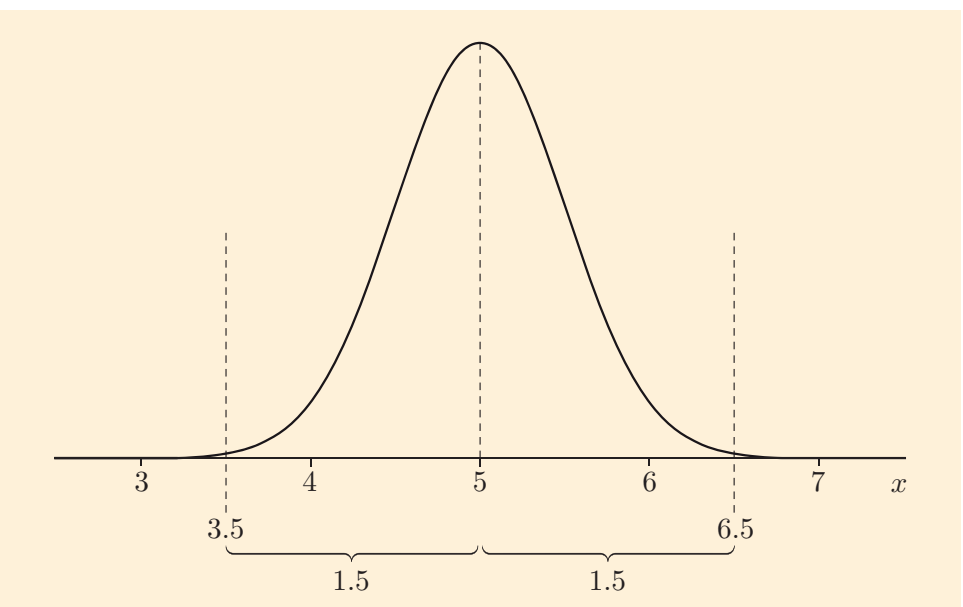


Figure 16 Investigating the spread of a normal distribution

The dashed lines in Figure 16 indicate that almost all of the distribution is contained between $x = 3.5$ and $x = 6.5$ (i.e. within 5 ± 1.5). This means that $3\sigma \simeq 1.5$, so $\sigma \simeq 0.5$.

In summary, the normal distribution plotted in Figure 15 is approximately the normal distribution with mean $\mu = 5$ and standard deviation $\sigma = 0.5$.

Activity 14 *Approximate values for the mean and standard deviation*

For each of the normal distributions shown in the parts of this activity, find approximate values for the mean and standard deviation, using the method described above.

(a)

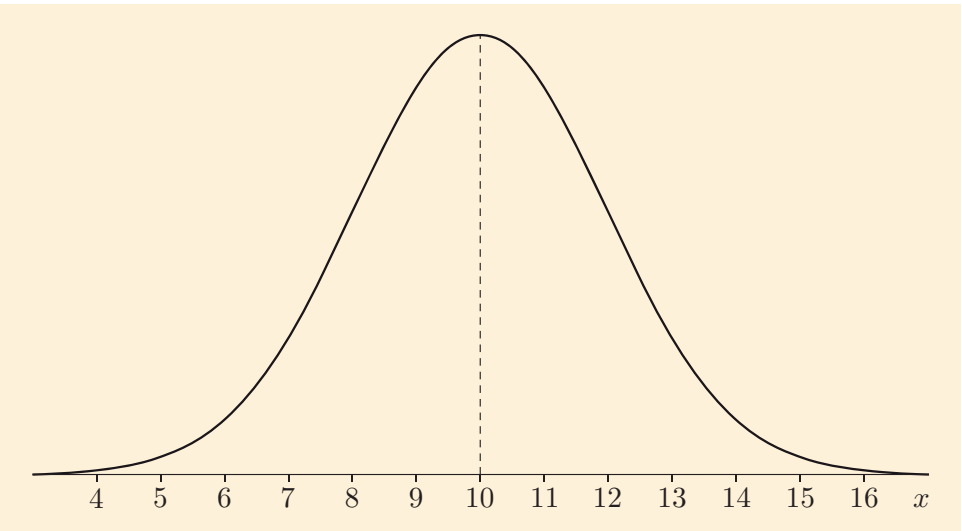


Figure 17 Another normal distribution

(b)

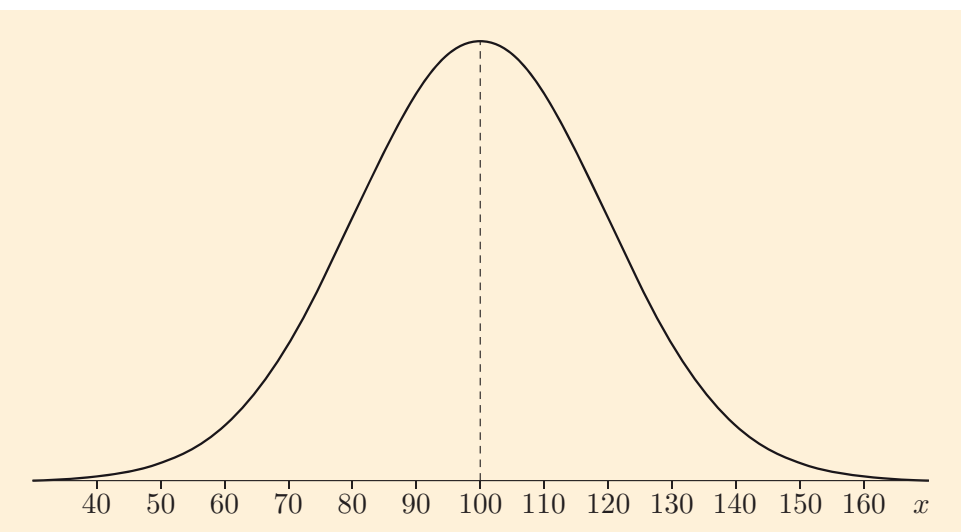


Figure 18 Yet another normal distribution

(c)

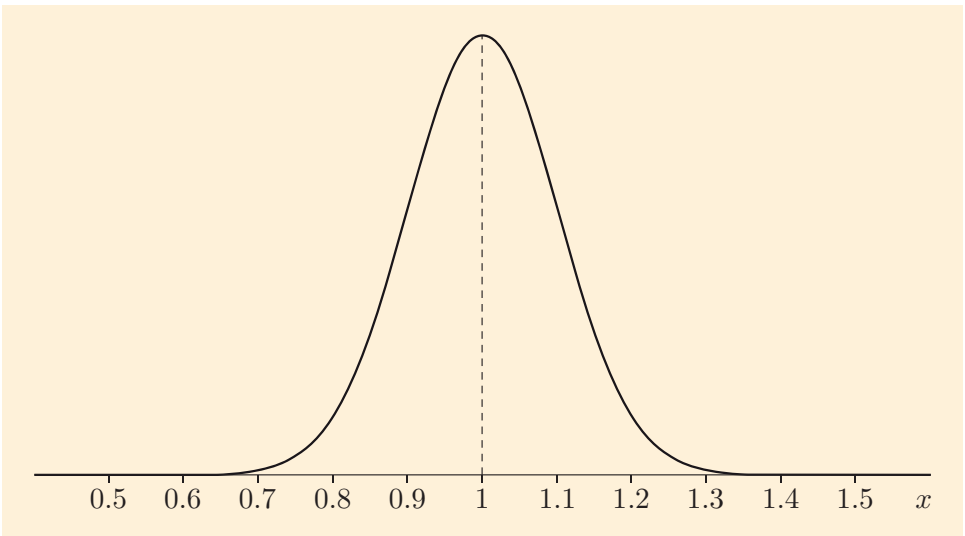
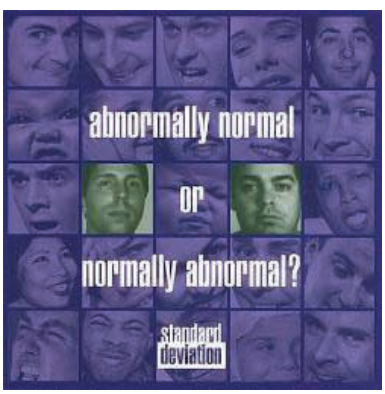


Figure 19 And yet one more normal distribution

Conversely, knowing the mean and standard deviation of a normal distribution enables us to make a rough sketch of the distribution. Any sketch of a normal distribution will show a symmetric and bell-shaped curve. More specifically, the distribution must be symmetric about the mean. In addition, almost the whole of the distribution must be contained within plus or minus three standard deviations of the mean.



The 1997 music CD 'abnormally normal or normally abnormal?' by the band 'standard deviation'.

Example 3 *Sketching a normal distribution*

The normal distribution of a variable x has mean $\mu = 15$ and standard deviation $\sigma = 3$. To sketch this distribution, draw a symmetric, bell-shaped curve centred on the value of μ, which in this case is 15. The standard deviation is $\sigma = 3$, so that $3\sigma = 9$. We therefore know that just about all the distribution is contained within 15 ± 9 (i.e. lies between $15 - 9 = 6$ and $15 + 9 = 24$). A sketch of the distribution can therefore be drawn and should resemble Figure 20.

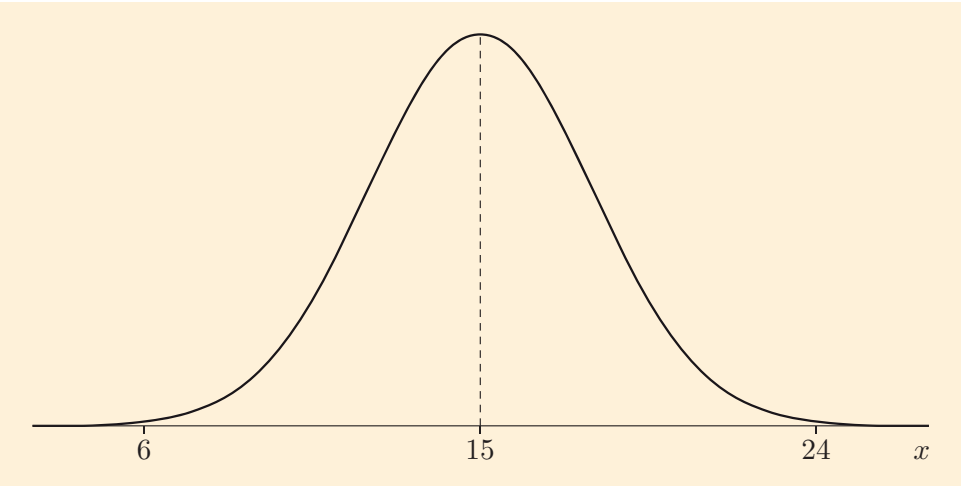


Figure 20 The normal distribution with $\mu = 15$ and $\sigma = 3$

The scale that is used for the horizontal axis certainly affects the shape of the normal distribution, as demonstrated by Figure 21. The important thing, though, is that the information conveyed by the sketch remains exactly the same.

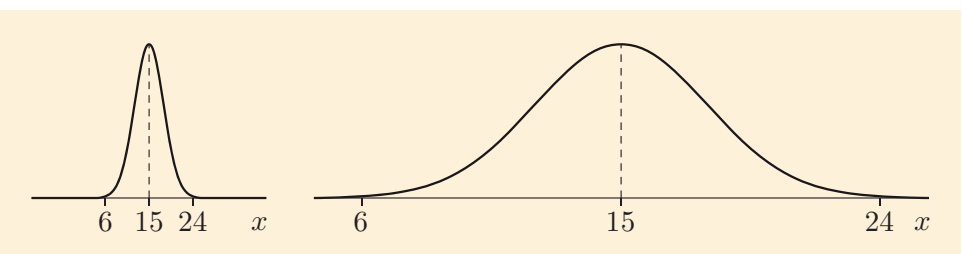


Figure 21 The same normal distribution plotted on different horizontal scales

Also, for the aspects we are investigating, the height of the distribution does not really matter; all the information we require about the relationship between the distribution and its mean and standard deviation is provided by the scale on the horizontal axis. For this reason there is no need to bother with a vertical scale at all.

Activity 15 *Sketching a normal distribution*

Sketch the following distributions:

- The normal distribution of a variable x with mean 1000 and standard deviation 100.
- The normal distribution of a variable x with mean 2 and standard deviation 0.25.

Activity 15 demonstrates that it always makes sense to think of the horizontal axis of a normal distribution in terms of the number of standard deviations of the variable away from the mean. This is illustrated in Figure 22, which is an important picture in understanding the normal distribution. Notice how the horizontal scale is marked off using μ and σ.

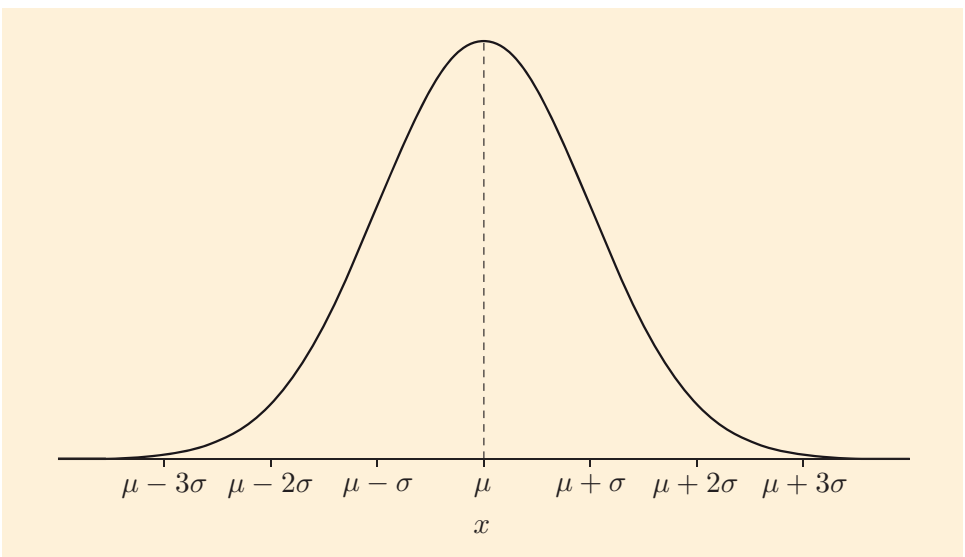


Figure 22 The normal distribution with its scale marked in terms of μ and σ

You have now covered the material related to Screencast 2 for Unit 7 (see the M140 website).

3.3 The standard normal distribution

We can go one step further than that represented by Figure 22 (Subsection 3.2) and think of all normal distributions in terms of one special normal distribution. This special normal distribution has mean zero and standard deviation one, and is called the **standard normal distribution**. It looks like Figure 23. Figure 23, in turn, looks like Figure 22 with μ and σ in the labels on the horizontal axis replaced by 0 and 1, respectively: so, $\mu - 3\sigma$ has become $0 - (3 \times 1) = -3$, $\mu - 2\sigma$ has become $0 - (2 \times 1) = -2$, and so on.

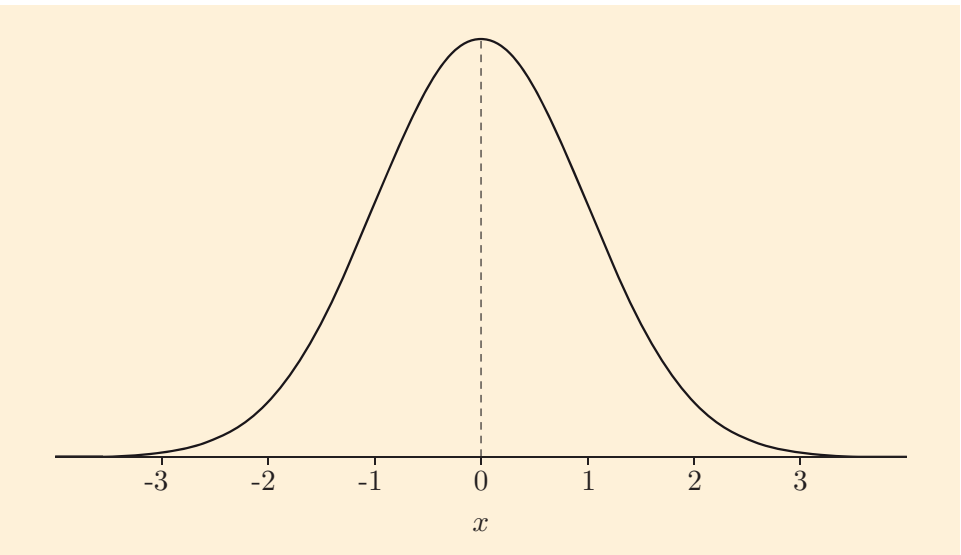


Figure 23 The standard normal distribution

The standard normal distribution

The standard normal distribution is the particular normal distribution that has mean $\mu = 0$ and standard deviation $\sigma = 1$.

It turns out that we can *transform* all normal distributions to the standard normal distribution.

Example 4 *Transforming to the standard normal distribution*

The normal distribution of a variable x with mean $\mu = 10$ and standard deviation $\sigma = 2$ is illustrated in Figure 24.

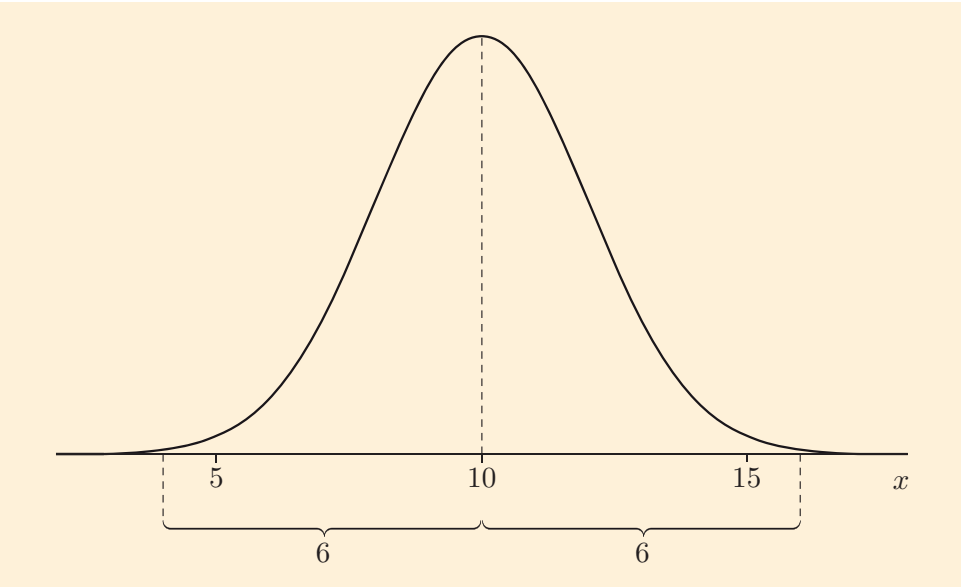


Figure 24 The normal distribution with $\mu = 10$ and $\sigma = 2$

First, we can shift the whole of the distribution to the left so that the mode occurs at zero just by subtracting 10 from each value of x. This is shown in Figure 25. It changes the *location* of the distribution but leaves the *spread* unchanged.

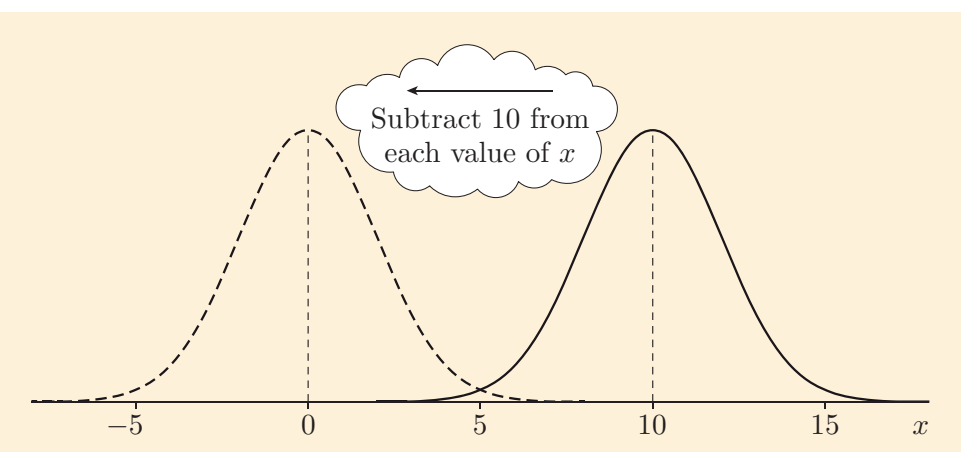


Figure 25 Shifting the distribution of x

The dashed curve in Figure 25 is now a new normal distribution with mean zero and standard deviation 2. This new distribution is the distribution of the variable v, say, where $v = x - 10$. The normal distribution of v differs from the standard normal distribution only by having standard deviation 2 rather than 1. However, if we now think of the

horizontal axis in terms of the number of standard deviations of v away from the mean, then we obtain Figure 26.

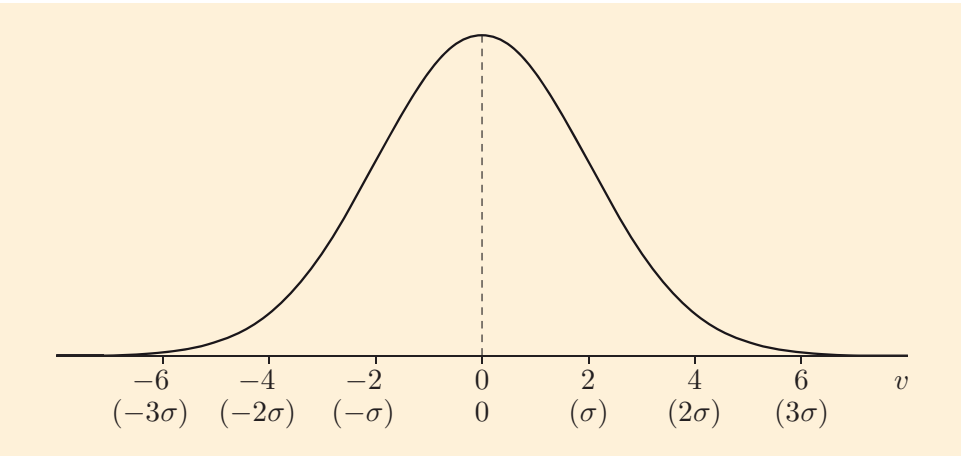


Figure 26 The normal distribution of $v = x - 10$ with mean 0 and standard deviation 2

Then, dividing every value of v by the standard deviation 2 gives the distribution of $v/2$. This distribution, shown in Figure 27, is the standard normal distribution, as required.

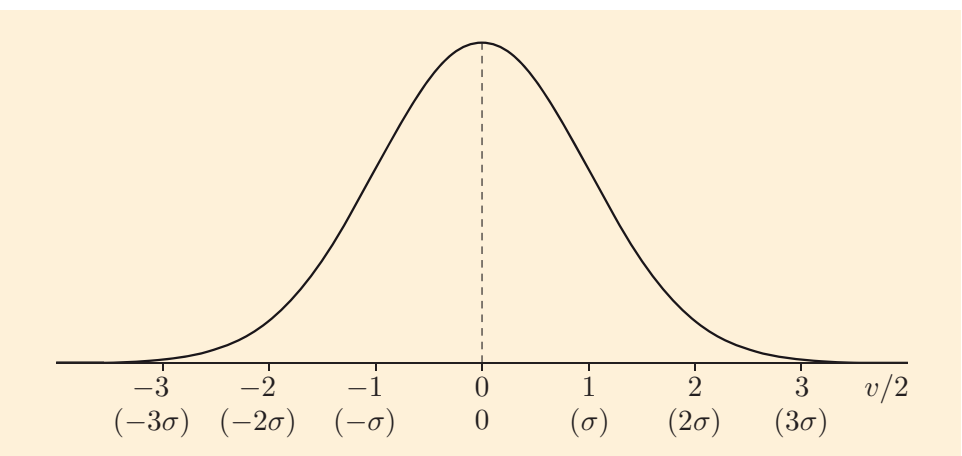


Figure 27 The normal distribution of $v/2$ with mean 0 and standard deviation 1

We have shown that if the variable x has a normal distribution with mean 10 and standard deviation 2, then the variable $v/2 = (x - 10)/2$ has the standard normal distribution.

Example 4 is a specific example of the following general result. If we start with a normal distribution for x, with mean μ and standard deviation σ, then:

- By subtracting μ from each value of x we obtain the distribution of $v = x - \mu$. This distribution is normal with mean zero and standard deviation σ.
- By then dividing each value of v by σ we obtain the variable $z = v/\sigma$, which has the standard normal distribution.

Combining the formulas for z and v, we find that

$$z = \frac{x - \mu}{\sigma}.$$

Transforming a normal distribution to the standard normal distribution

If a variable x has a normal distribution with mean μ and standard deviation σ, then the variable

$$z = \frac{x - \mu}{\sigma}$$

has the standard normal distribution.

Activity 16 *Transforming some particular normal distributions*

For the normal distributions with the following values of μ and σ, write down the appropriate formula to transform the variable x to the variable z that follows the standard normal distribution.

(a) $\mu = 10$, $\sigma = 2$ (b) $\mu = 100$, $\sigma = 20$ (c) $\mu = 1$, $\sigma = 0.1$

Activity 17 *Transforming the distribution of Scottish men's heights*

(a) Assume that the population distribution of Scottish men's heights h (in metres) is normal with mean $\mu = 1.75$ and standard deviation $\sigma = 0.07$. Write down the formula for z which transforms each value of the variable h to the number of standard deviations from its mean.

(b) Calculate the value of z corresponding to each of the following values of h (in metres). In each case, interpret your answer by completing a sentence of the form 'So a height of *** metres is *** standard deviations *** the mean height of *** metres'.

$h = 1.96$; $h = 1.61$; $h = 1.785$.

Importance of the standard normal distribution

The development in this subsection implies that by describing every normal distribution in terms of z, the number of standard deviations by which the variable differs from its mean, we can think of all normal distributions in terms of just one distribution: the standard normal distribution.

Wall space increased at the Statistics Art Gallery when it became clear that only one picture was required in the 'normal distribution' collection.

You have now covered the material needed for Subsection 7.2 of the Computer Book.

You have also now covered the material related to Screencast 3 for Unit 7 (see the M140 website).

Exercises on Section 3

Exercise 6 *Approximating the mean and standard deviation*

Find approximate values for the mean and standard deviation of the normal distribution shown below.

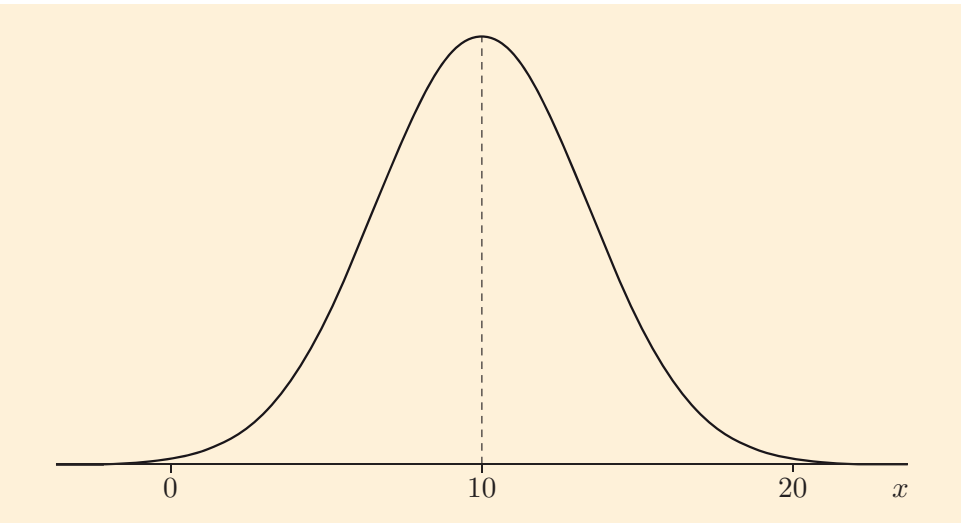


Figure 28 Yet again, a normal distribution

Exercise 7 *Approximating another mean and standard deviation*

Find approximate values for the mean and standard deviation of the normal distribution shown below.

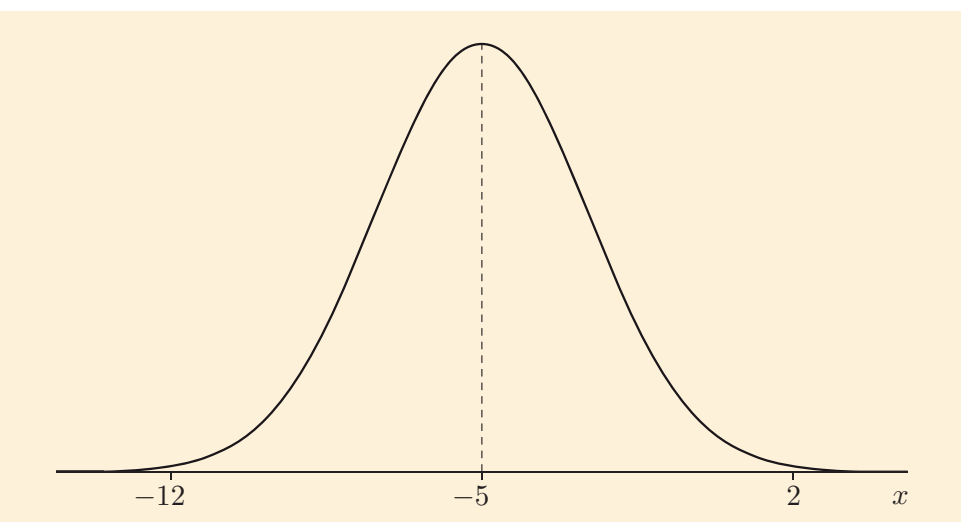


Figure 29 And one more time, another normal distribution

Exercise 8 *Sketching a normal distribution*

The normal distribution of a variable x has mean -1 and standard deviation 1. Sketch the distribution.

Exercise 9 *Sketching another normal distribution*

The normal distribution of a variable x has mean 4 and standard deviation 4. Sketch the distribution.

Exercise 10 *Obtaining z for a normal distribution*

Write down the appropriate formula to transform the variable x to the variable z that follows the standard normal distribution when

(a) x has the normal distribution with mean 6 and standard deviation 3.3;

(b) x has the normal distribution with mean -6 and standard deviation 2.

Exercise 11 *Calculating z from x*

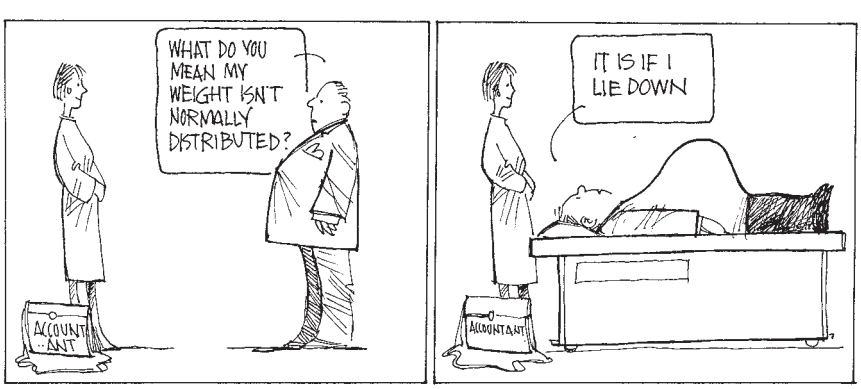
Assume that x follows the normal distribution with mean $\mu = 2$ and standard deviation $\sigma = 10$. Write down the appropriate formula for z which transforms the variable x to the number of standard deviations from its mean. Calculate the value of z corresponding to $x = 3$.

Exercise 12 *Calculating z from x for another normal distribution*

Assume that x follows the normal distribution with mean $\mu = -1$ and standard deviation $\sigma = 0.5$. Write down the appropriate formula for z which transforms the variable x to the number of standard deviations from its mean. Calculate the value of z corresponding to $x = 0$.

4 Sampling distributions re-revisited

We now take a closer look at the sampling distributions of the sample mean that you met in Section 2. As we said there, provided the sample size is sufficiently large (roughly speaking, greater than 25), these sampling distributions are approximately normal. Thus the ideas discussed in Section 3, which apply to *all* normal distributions, apply (approximately) to these sampling distributions as well. These ideas will enable us to find a suitable test statistic to use for testing some of the hypotheses we are interested in for the BCS survey.



We begin by examining the relationship between sampling distributions of the mean and the original population distribution in a little more detail.

Activity 18 *Means of distributions of sample means*

(a) Consider again the population distribution of MS221 examination marks which you met in Section 2. In fact, this population distribution has mean $\mu = 66$ and standard deviation $\sigma = 22$. Figure 30 shows the sampling distributions of the mean for various sample sizes. (Figure 30 is similar to Figure 6 but for some different values of n.)

What do you notice about the means of these sampling distributions compared with the population mean?

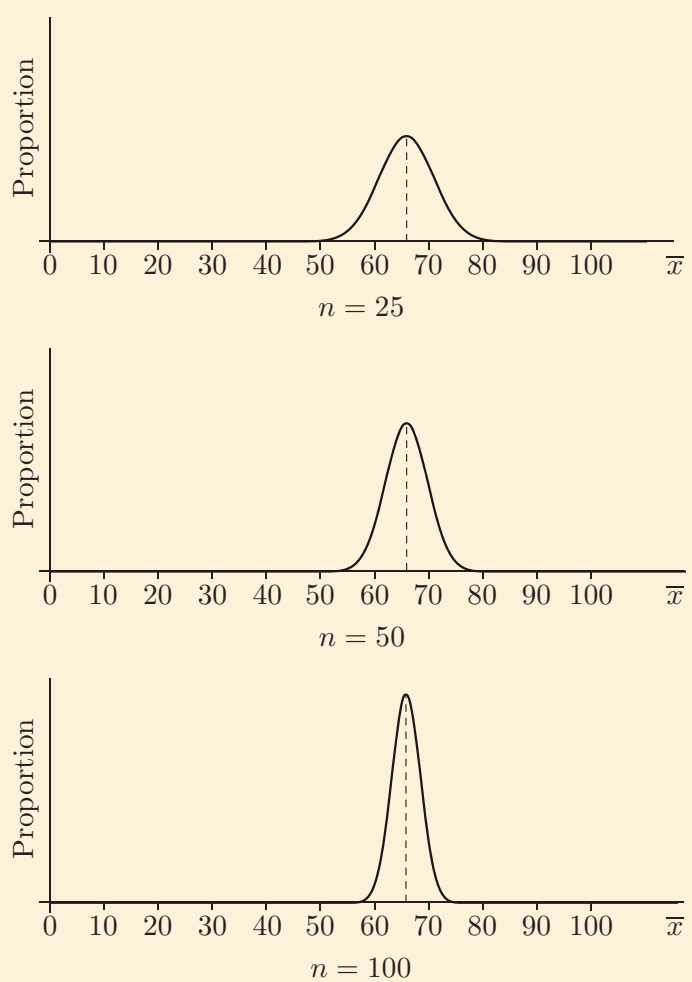


Figure 30 Sampling distributions of the mean for samples of size n from the population of exam marks

(b) Consider again the population distribution of full-time employees' earnings which you met in Example 1, in Section 2. This population distribution has mean $\mu = 491$ and standard deviation $\sigma = 283$ (in £). Figure 31 shows again the sampling distributions of the mean for

various sample sizes. (Figure 31 is similar to Figure 8 but for different values of n.)

What do you notice about the means of these sampling distributions compared with the population mean?

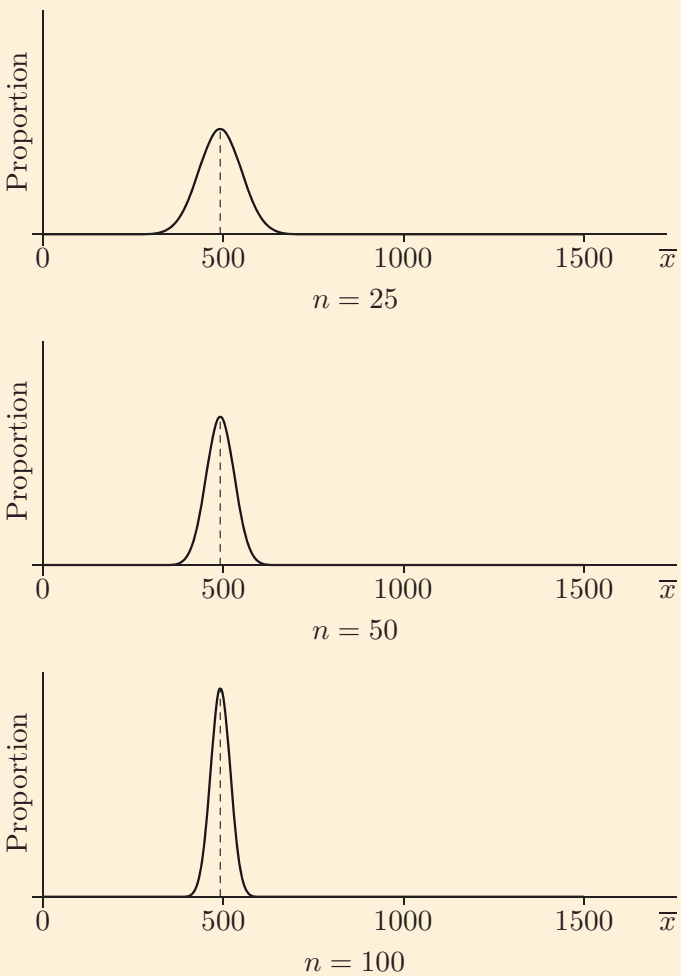


Figure 31 Sampling distributions of the mean for samples of various sizes from the population of employees' earnings

The conclusions of Activity 18 hold more generally so that *whatever* the population distribution (no matter what shape) and *whatever* the sample size (no matter how small), the mean of the sampling distribution is always equal to the population mean μ.

Now let us take a closer look at the *spread* of the sampling distributions.

Activity 19 *Standard deviations of distributions of sample means (1)*

(a) Consider again the sampling distributions of the mean for MS221 examination marks that are shown in Figure 30. What do you notice about the standard deviations of these sampling distributions?

(b) Consider again the sampling distributions of the mean for full-time employees' earnings that are shown in Figure 31. What do you notice about the standard deviations of these sampling distributions?

In fact it can be shown that for population standard deviation σ, the standard deviation of the sampling distribution of the mean for samples of size n is $\sigma/\sqrt{n}$.

Activity 20 *Standard deviations of distributions of sample means (2)*

The population distribution of examination marks has standard deviation $\sigma = 22$. Use the formula to find the standard deviation of the sampling distribution of the mean for samples of size

(a) 25; (b) 50; (c) 100.

Both the formula $\sigma/\sqrt{n}$ and the calculations in Activity 20 confirm that the standard deviation of the sampling distribution of the mean does decrease as n increases, as was suggested in Activity 19. What is not so clear, and is perhaps unexpected, is the precise way in which the standard deviation of the sampling distribution of the mean depends on n – through its square root.

The expression 'standard deviation of the sampling distribution of the mean' is a bit of a mouthful. It is often referred to as the **standard error of the mean** for samples of size n, or sometimes just the **standard error** for short, in which case it can be abbreviated to the symbol SE. Using this abbreviation, we obtain the formula $\text{SE} = \sigma/\sqrt{n}$, which is easier to remember.

The terminology 'standard *error*' is related to the notion of 'sampling *error*', which you met in Subsection 4.1 of Unit 4.

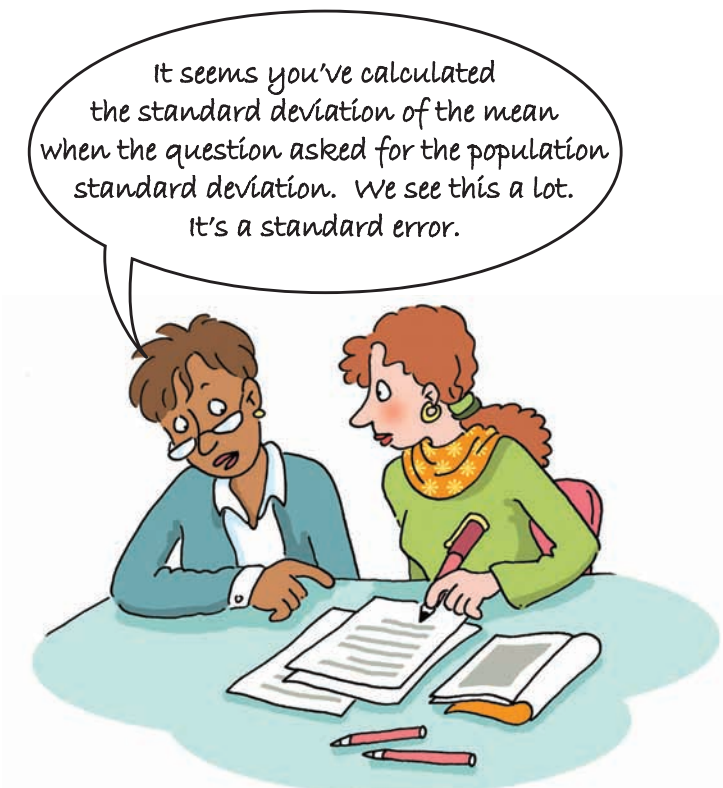
The above result holds generally for *all* sampling distributions, no matter *what* the population distribution and no matter *what* sample size is involved. So there is a very precise relationship between sampling distributions and the population distribution. It can be summarised as follows.

Mean and standard deviation of the sampling distribution of the mean

- The mean of the sampling distribution is equal to μ, the population mean.
- The standard deviation of the sampling distribution is called the standard error of the mean. It is given by

$$\text{SE} = \frac{\sigma}{\sqrt{n}},$$

where n is the sample size and σ is the population standard deviation.



The relationship between sampling distributions and the population distribution is particularly useful when the sample size is large and the sampling distribution is approximately normal. In practice, we usually have very little information about the population distribution itself. Indeed we often have only a sample of data on which to base our analysis; there is *no* other information about the population. Yet many techniques of statistical inference require us to make some assumptions about the population distribution.

The advantage of working with large samples is that, no matter what shape the population distribution is, the sampling distribution of the mean for samples of size n will *always* be more or less normal. Moreover, we

know that the mean of this sampling distribution is equal to the population mean, μ, and the standard deviation is the standard error, given by $\text{SE} = \sigma/\sqrt{n}$, where σ is the population standard deviation. This is summarised below.

Approximate normality of the sampling distribution of the mean

If n is large, no matter what shape the population distribution is, the sampling distribution of the mean for samples of size n will be approximately normal with mean equal to the population mean, μ, and standard deviation equal to the standard error, $\text{SE} = \sigma/\sqrt{n}$.

(This important result is often called the *central limit theorem*.)

Activity 21 *Approximate distribution of ball bearing diameters*

The population distribution of the diameters of ball bearings produced by a particular manufacturer has mean $\mu = 2\,\text{mm}$ and standard deviation $\sigma = 0.01\,\text{mm}$. Find the standard deviation of the sampling distribution of the mean for samples of 25 such ball bearings. Hence give the approximate distribution of the mean diameter of ball bearings in samples of size 25.

What this implies is that we can base our analysis on the relationship between the sample data and the sampling distribution of the mean. Thus we infer back from the evidence provided by the sample data to the sampling distribution. Then our knowledge of the links between this sampling distribution and the population distribution allows us to draw conclusions about the population mean. This is a very important strategy in statistics.

The new hypothesis test, the z-test, is based on just this principle and will be fully discussed in Section 5. As you now know, the sampling distribution of the mean, $\overline{x}$, for large samples of size n is approximately normal with mean μ and standard deviation $\text{SE} = \sigma/\sqrt{n}$. As with any normal distribution, we can transform this normal sampling distribution into the standard normal distribution. This means that the distribution of the variable

$$z = \frac{\overline{x} - \mu}{\text{SE}}$$

is the standard normal distribution (with mean zero and standard deviation one). There is a strong connection between this result and the z-test to follow.

You have now covered the material related to Screencast 4 for Unit 7 (see the M140 website).

Exercises on Section 4

Exercise 13 *Standard deviations of the mean as sample size changes*

The population distribution of full-time employees' earnings has standard deviation $\sigma = 283$. Find the standard deviation of the sampling distribution of the mean for samples of size

(a) 9; (b) 25; (c) 100.

Exercise 14 *Standard deviations of another mean*

The population distribution of a certain quantity has standard deviation $\sigma = 3.6$. Find the standard deviation of the sampling distribution of the mean for samples of size

(a) 4; (b) 19; (c) 300.

Exercise 15 *Standard deviation of the average content of water bottles*

The population distribution of the amount of water contained in a nominally one-litre bottle from a certain manufacturer has mean $\mu = 1.01$ litres and standard deviation $\sigma = 0.01$ litres. Find the standard deviation of the sampling distribution of the mean for samples of 40 such bottles. Hence give the approximate distribution of the mean amount of water contained in samples of 40 one-litre bottles from this manufacturer.

5 The one-sample z-test

Unless we need to distinguish the one-sample z-test from the two-sample z-test that will be developed in Section 6, we often omit the phrase 'one-sample'.

In this section we shall develop a new hypothesis test, the **one-sample z-test**. The hypotheses are concerned with the mean, μ, of the population from which the sample is selected. We shall suppose that a particular value, say A, is of special interest as a potential value for μ. The null hypothesis is

$$H_0\colon\ \mu = A,$$

and the alternative hypothesis is

$$H_1\colon\ \mu \neq A.$$

One-sided alternative hypotheses will be discussed in Unit 10.

Alternative hypotheses of this form are often called **two-sided alternative hypotheses**. This is because they include both $\mu < A$ and $\mu > A$.

The above is the first of the four stages of hypothesis testing that you were introduced to at the start of Section 4 of Unit 6. In abbreviated form, these are:

1. Set up the hypotheses that we wish to test.
2. Determine the sampling distribution of a test statistic under the assumption that the null hypothesis is true.
3. Ascertain how unlikely the observed value of the test statistic is on the basis of the sampling distribution.
4. If the test statistic turns out to have a very unlikely value, then either:
   - a very unusual event has happened, or
   - the sample has provided evidence against the correctness of the null hypothesis.

To develop ideas in the current context, we first consider the simpler case where the population standard deviation is assumed to be known, and in Subsection 5.2 we consider the more realistic case where it is unknown. The tests that are developed make use of the results presented in Section 4 about the sampling distribution of the sample mean.

5.1 The z-test with the standard deviation assumed to be known

To describe the z-test we will use a simple (constructed) example.

Example 5 *Has a new method of teaching made a difference?*

For many years a teacher has been using the same method of teaching children to read. The scores the children obtain on a reading test have a mean of 54.6 and a standard deviation of 8.3. These values will be taken to be the population mean and the population standard deviation under the old method of teaching. The teacher tries a new method with her current class of 34 children, and their average score on the reading test is 58.1. She wants to test whether random variation underlies the difference between the average of this class (58.1) and the long-term average of previous classes (54.6), or whether there is a genuine difference.

The null and alternative hypotheses are:

H_0: The old method and new method of teaching children to read are equally effective.

H_1: The old method and new method of teaching children to read differ in their effectiveness.

If μ denotes the mean reading score of children taught by the new method, we can recast these hypotheses as

H_0: $\mu = 54.6$

H_1: $\mu \neq 54.6$.

The sample mean, $\overline{x}$, is based on the performances of $n = 34$ children. Hence, its sampling distribution is approximately normal, as a sample size of 34 is quite large. Moreover, as shown in Section 4:

- the mean of the sampling distribution of $\overline{x}$ is equal to μ
- the standard deviation of the sampling distribution of $\overline{x}$ (i.e. the standard error of $\overline{x}$) is equal to $\sigma/\sqrt{n}$.

Now, to perform a hypothesis test based on $\overline{x}$, the sampling distribution under which we calculate probabilities is the sampling distribution of $\overline{x}$ *assuming that the null hypothesis, H_0, is true.*

In the present case, if H_0 is true, then $\mu = 54.6$ and the distribution of $\overline{x}$ is approximately normal with mean 54.6 and standard deviation $\sigma/\sqrt{n}$. We know that n equals 34 but need to know the value of σ. For this example, we shall assume that the population standard deviation of scores with the new method is the same as with the old method, so $\sigma = 8.3$. All told,

$$A = 54.6, \quad \overline{x} = 58.1, \quad n = 34, \quad \sigma = 8.3.$$

Now, from the end of Section 4, if the sampling distribution of $\overline{x}$ is approximately normal with mean μ and standard deviation $\text{SE} = \sigma/\sqrt{n}$, then the distribution of the variable

$$z = \frac{\overline{x} - \mu}{\text{SE}}$$

is (approximately) the standard normal distribution (with mean zero and standard deviation one). Thus, if H_0 is true, so that $\mu = A = 54.6$, the distribution of the variable

$$z = \frac{\overline{x} - 54.6}{\text{SE}}$$

is (approximately) the standard normal distribution.

The variable z is the test statistic for the z-test. Its numerical value in this example is

$$z = \frac{\overline{x} - A}{\text{SE}} = \frac{58.1 - 54.6}{8.3/\sqrt{34}} \simeq 2.46.$$

The main result that we have obtained so far is summarised in the following box.

Test statistic and its sampling distribution when H_0 is true and σ is assumed known

For a one-sample z-test, when H_0: $\mu = A$ is true, the test statistic,

$$z = \frac{\overline{x} - A}{\text{SE}},$$

follows (approximately) the standard normal distribution, where

$$\text{SE} = \sigma/\sqrt{n}.$$

Activity 22 *Value of z*

Calculate the value of the test statistic z for the test of

H_0: $\mu = 120$
H_1: $\mu \neq 120$,

when $n = 100$, $\overline{x} = 112$, and $\sigma = 15$.

Critical values and critical regions

If the null hypothesis, H_0, is true, then z should follow the standard normal distribution. This distribution has a mean of 0, so if the value of z given by our data was very large in size (positive or negative), it would suggest that H_0 is false. The idea, then, is to reject H_0 if the observed value of z is 'too extreme' and therefore unlikely. Notice that 'too extreme' covers both large positive values and large negative values, in line with H_1, which specifies $\mu \neq A$, 'in either direction' away from A. If we cannot believe that the observed z is an observation from a standard normal distribution, then we cannot believe H_0.

We have calculated the value of the test statistic z that is given by our data. Suppose now that the test is to be performed at the 5% significance level. As discussed in Subsection 4.1 of Unit 6, H_0 will be rejected at the 5% significance level if z is in the most extreme 5% of values under the sampling distribution that applies if H_0 is true. This 'most extreme' region is the **critical region** of the test. (In this case it is the critical region at the 5% significance level.) Because of the discussion in the previous paragraph, the critical region consists of two parts: one part comprises the most extremely high 2.5% of values under the standard normal distribution, and the other part comprises the most extremely low 2.5% of values under the standard normal distribution.

The values defining the 'inner ends' of the critical region are the **critical values**. The critical values for the z-test at the 5% significance level are 1.96 and -1.96. Figure 32 shows the critical values and critical region pictorially.

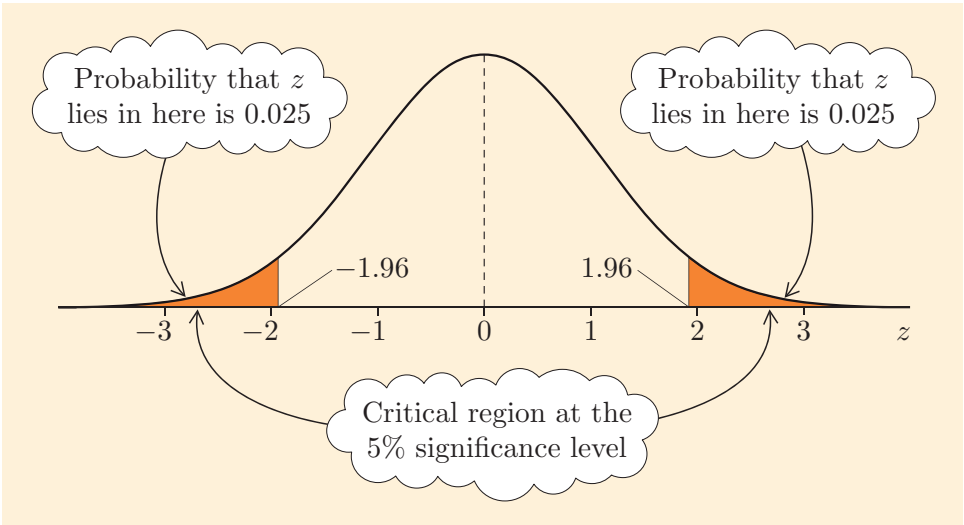


Figure 32 The standard normal distribution with the critical region and critical values (1.96 and −1.96) shown for a test at the 5% significance level

Instead of using the 5% significance level for the hypothesis test, we might want to perform the test at the more stringent 1% significance level. To do this, all that changes is the values of the critical values and hence the critical region. The critical values become 2.58 and −2.58, and the critical region is rather smaller: see Figure 33.

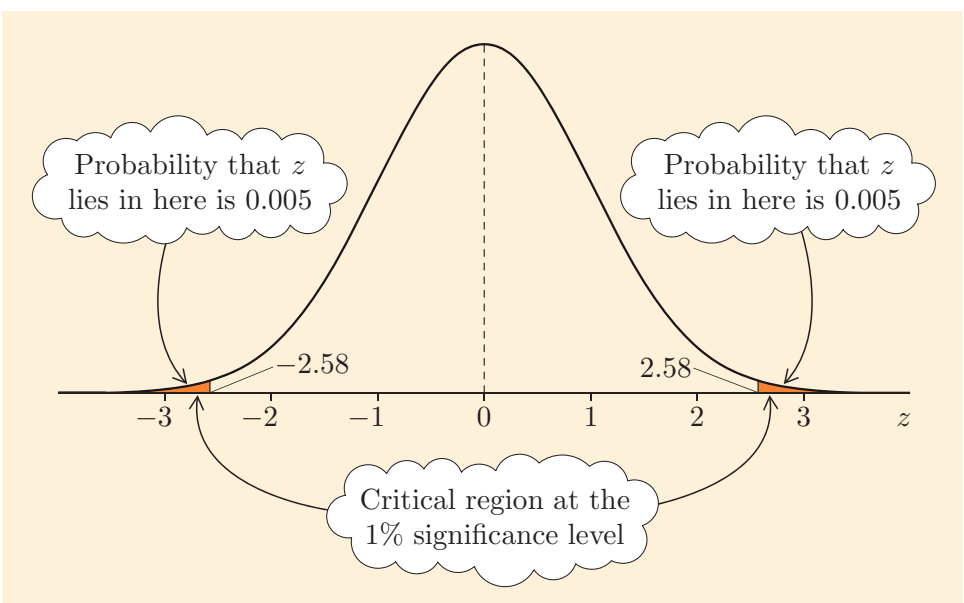


Figure 33 The standard normal distribution with the critical region and critical values (2.58 and −2.58) shown for a test at the 1% significance level

The procedure to be followed to complete the z-test is as follows.

Completing the z-test

If $z \geq 2.58$ or $z \leq -2.58$, reject H_0 at the 1% significance level.

If $1.96 \leq z < 2.58$ or $-2.58 < z \leq -1.96$, reject H_0 at the 5% significance level but not at the 1% significance level.

If $-1.96 < z < 1.96$, do not reject H_0 at the 5% significance level.

Activity 23 *Where is z in marginal circumstances?*

On a sketch of the standard normal distribution, show where the value of z must lie in the marginal case where H_0 is rejected at the 5% significance level but not at the 1% significance level.

As noted in Subsection 5.1 of Unit 6, we conclude that there is strong evidence against the null hypothesis if we reject H_0 at the 1% significance level. If we reject it at the 5% significance level but not the 1% level, we conclude that there is moderate (but not strong) evidence against the null hypothesis. If we do not reject H_0 at the 5% significance level, we have, in the words of Subsection 5.1 of Unit 6, either 'little' or 'weak' evidence against H_0.

We will use 'strong' whenever we reject H_0 at the 1% significance level in this unit; the evidence might in fact be 'very strong', but we will not be testing H_0 at the 0.1% level.

Example 6 *Completing the z-test started in Example 5*

In Example 5, the test statistic (the data z-value) takes the value 2.46. This value exceeds 1.96 but not 2.58. We conclude that there is moderate evidence that the old and new methods are not equally effective at teaching children to read.

As the new method gave an average score of 58.1, while the average under the old method was 54.6, and a higher test score means better reading ability, there is moderate evidence that the new method is better than the old method.

Activity 24 *A z-test in manufacturing*

A firm is engaged in putting finishes on work surfaces for kitchen manufacturers. Previously, the work was done in very large batches, so the time spent setting up the machine did not affect production too much. However, with a change in the pattern of demand the batch size has had to be considerably reduced, so the time spent setting the machine to different specifications is becoming more important.

Last year the manufacturing manager found that the machine setting had been changed very many times and the mean time taken for a change was 26.1 minutes. The operators suggested a way in which the set-up time might be reduced, but the manager was unconvinced and feared that the set-up time might actually be increased. Nevertheless, it was agreed to try

out this new method for two weeks. A z-test would then be performed to examine whether or not the mean time for setting up under the new method differs from the mean time taken last year.

In the two-week testing period, the machine was reset on 53 occasions, taking a mean time of 20.9 minutes.

(a) What are the appropriate null and alternative hypotheses?

(b) Give the values of A, $\overline{x}$ and n.

(c) Assume that the standard deviation, σ, equals 12.3. Calculate the value of the test statistic.

(d) Is the null hypothesis rejected at the 5% significance level? Is it rejected at the 1% significance level?

(e) What do you conclude from the hypothesis test?

5.2 The z-test with unknown standard deviation

In Subsection 5.1, we developed one-sample z-tests under the assumption that σ is known. Now σ is the standard deviation of the population from which the sample data are drawn. Typically its actual value will not be known, but if we have a large sample then the sample standard deviation, s, provides a good estimate of σ. Moreover, provided the sample size is large, the one-sample z-test can be performed with σ replaced by s. Specifically, we calculate the *estimated standard error* (ESE) of $\overline{x}$,

Compare formula for ESE with SE $= \sigma/\sqrt{n}$.

$$\text{ESE} = \frac{s}{\sqrt{n}},$$

and put

$$z = \frac{\overline{x} - A}{\text{ESE}}.$$

You might be slightly disquieted by the bald assertion that, for large samples, replacing SE by its estimated value ESE makes no difference to the (approximate) standard normal distribution of $(\overline{x} - A)/\text{SE}$. After all, ESE is not the correct quantity to divide by; SE is. It is the assumption of a large sample that saves the day. In Unit 10 we give tests for small samples (t-tests) which take the difference between ESE and SE into account. Differences between those tests and z-tests are small when the sample size is above about 25.

At the end of Section 2, it was asserted that the sampling distribution of the mean will always be approximately normal for sample sizes greater than 25. It was also argued that the sampling distribution of the mean will actually be approximately normal for sample sizes (much) smaller than $n = 25$ for many population distributions. In that sense, the notion of $n = 25$ being large enough errs on the 'careful' side. When SE is replaced by its estimated value (ESE), however, a sample size of 25 is only just enough for a z-test to be usable. We will continue to use this 'rule of

thumb', but $n = 25$ is no longer a 'generous' value – many would prefer to use z-tests only for samples that are a bit larger than that.

What is a large enough sample for a z-test?

As a rough guide you can assume that, whatever the population distribution, for sample sizes of at least 25, the z-test is applicable.

As this jolly logo shows, ESE also stands for 'Exceptional Student Education' ... an educational program in schools in Citrus County, Florida, USA

The next two boxes lay out the full requirements and procedure for the one-sample z-test. They cover both the cases where σ is known and where it must be estimated. The first box gives the key pieces of information that you should pick out for a z-test when you are reading details about a survey or experiment.

Key values for a one-sample z-test

The information you need to know for a one-sample z-test is:

- the hypothesised population mean (A) under the null hypothesis
- the sample mean ($\overline{x}$)
- the sample size (n)
- the population standard deviation (σ), or a good estimate of σ.

Procedure: the one-sample z-test

1. Set up the null and alternative hypotheses,

   H_0: $\mu = A$
   H_1: $\mu \neq A$,

   where μ is the population mean.

2. Calculate the test statistic, z:

   - If the population standard deviation (σ) is known,

     $$z = \frac{\overline{x} - A}{\text{SE}}, \quad \text{where SE} = \frac{\sigma}{\sqrt{n}}.$$

   - If σ is unknown but the sample size (n) is 25 or more,

     $$z = \frac{\overline{x} - A}{\text{ESE}}, \quad \text{where ESE} = \frac{s}{\sqrt{n}}.$$

   Here $\overline{x}$ is the sample mean and s is the standard deviation of the sample. SE is the standard error of the mean and ESE is the estimated standard error.

3. Compare z with the appropriate critical values, which are 1.96 and -1.96 at the 5% significance level and 2.58 and -2.58 at the 1% significance level.

   - If $z \geq 2.58$ or $z \leq -2.58$, then H_0 is rejected at the 1% significance level.
   - If $1.96 \leq z < 2.58$ or $-2.58 < z \leq -1.96$, then H_0 is rejected at the 5% significance level but not at the 1% significance level.
   - If $-1.96 < z < 1.96$, then H_0 is not rejected at the 5% significance level.

4. State the conclusions that can be drawn from the test.

We are now in a position to start answering some of the questions we asked about the BCS survey data in Subsection 1.3. The investigation illustrates use of the one-sample z-test when σ is unknown.

Example 7 *Reading scores of 7-year-old children in BCS survey*

In a question that was posed at the end of Subsection 1.3, we asked whether the sample of children from the BCS survey in 2004–2005 could be considered to have come from the population of children for whom the British Ability Scales reading score was developed. The overall population mean reading scores for British children are taken to be 96 for 7-year-old children. We wrote down the following null and alternative hypotheses:

H_0: For British children aged 7 in 2004–2005, the mean reading score is equal to 96

H_1: For British children aged 7 in 2004–2005, the mean reading score is not equal to 96.

We can recast these hypotheses as

H_0: $\mu = 96$

H_1: $\mu \neq 96$,

where μ is the population mean of the reading scores of all British 7-year-old children in 2004–2005. The data from the BCS concerning 7-year-old children are summarised in Table 3.

Table 3 Further summary statistics for data on reading scores of 7-year-old children

Sample size	Sample mean	Sample standard deviation
396	111.28	26.668

(This data is copyright and owned by the Economic and Social Data Service.)

Although σ is unknown, the sample size is considerably greater than 25, so the ESE may be used in calculating z. Thus the information required for the z-test is:

$$A = 96, \quad \overline{x} = 111.28, \quad n = 396, \quad s = 26.668.$$

We can now calculate the test statistic:

$$z = \frac{\overline{x} - A}{\text{ESE}} = \frac{\overline{x} - A}{s/\sqrt{n}} = \frac{111.28 - 96}{26.668/\sqrt{396}} \simeq 11.40.$$

Assuming that the null hypothesis, H_0, is true, 11.40 is a value from the standard normal distribution. However, 11.40 is much bigger than the 1% critical value of 2.58. Hence the z-test clearly rejects H_0 at the 1% level.

We conclude that there is strong evidence that the mean reading score for 7-year-old children in 2004–2005 is not equal to the overall mean reading score for 7-year-old children. At face value, this is a little surprising – there seems no obvious factor to cause a difference in reading ability between the 7-year-old children in the 2004–2005 BCS survey and the 7-year-olds in the population of British children for whom the reading test was originally developed. Given that the mean reading score for the 7-year-old children in the BCS survey is larger than the overall mean reading score, for some reason the BCS children seem to have performed rather better than expected (on average).

The following two activities provide you with practice in applying the z-test. The first one continues our investigation of the BCS data. It concerns the reading scores of 8-year-old children. The second concerns some data on earnings.

Activity 25 *Reading scores of 8-year-old children in BCS survey*

In the BCS investigation, the following results were obtained for 8-year-old children.

Table 4 Summary statistics for data on reading scores of 8-year-old children

Sample size	Sample mean	Sample standard deviation
283	126.92	27.711

(This data is copyright and owned by the Economic and Social Data Service.)

The overall mean reading score for 8-year-old children is 116.

Carry out a z-test to investigate whether the sample of 8-year-old children was selected from a population whose mean reading score is equal to the overall mean score for 8-year-old children. Comment on your result.

Activity 26 *Wages of female employees*

A random sample of 810 female local government clerical officers and assistants had a mean wage of £373.40 per week in 2011 with a standard deviation of £138.20. The overall mean weekly wage for female employees in 2011 was £381.50. (Source: *Annual Survey of Hours and Earnings*, 2011.) Investigate whether the mean weekly wage of female local government clerical officers and assistants differed from the overall mean weekly wage for female employees in 2011. Comment on your result.

You have now covered the material related to Screencast 5 for Unit 7 (see the M140 website).

Exercises on Section 5

Exercise 16 *Reading scores of 7-year-old girls in BCS survey*

In the BCS investigation, the following results were obtained for 7-year-old girls. (These results have been extracted from Table 2 in Subsection 1.3.)

Table 5 Summary statistics for data on reading scores of 7-year-old girls

Sample size	Sample mean	Sample standard deviation
190	113.42	25.464

(This data is copyright and owned by the Economic and Social Data Service.)

The overall mean reading score for 7-year-old children is 96.

Carry out a z-test to investigate whether the sample of 7-year-old girls was selected from a population whose mean reading score is equal to the overall mean score for 7-year-old children. Comment on your result.

Exercise 17 *Weight of pigs*

A random sample of 533 pigs of a certain breed that had been fed a special diet were weighed. They had a mean weight of 81.92 kg with a standard deviation of 15.65 kg. The mean weight of this breed of pig when fed the standard diet is 80 kg. Evaluate the evidence that the special diet changes the mean weight of this breed of pig.

Exercise 18 *An exciting exercise: paint drying*

A consumer magazine, when comparing various brands of paint, stated that the drying time of one particular brand was exactly four hours. The manufacturers of that paint were not particularly pleased with this as they believed the drying time for their paint was shorter. They organised a trial in which the paint was tested by a random sample of 40 customers, all of whom were decorating their living rooms. For this sample the mean drying time was found to be 3.80 hours and the standard deviation was 0.55 hours.

(a) Analyse the sample data to test whether the drying time given by the consumer magazine is correct.

(b) What reservations might there be about your conclusion?

In 2011/12, the internet (including one national newspaper) was abuzz with news of the forthcoming inaugural World Watching Paint Dry Championships to be held in Stoke-on-Trent in July 2012. Competitors were to each be given a one-metre square patch of freshly emulsioned wall at which to stare as it slowly dried. There were said to be 42 entrants, from the UK, USA, India and Hungary. Unfortunately, there is no evidence that the event actually took place.

6 The two-sample z-test

In this section we develop the **two-sample z-test**, which is used to analyse the difference in locations between two populations. There were plenty of examples of this raised in the context of the BCS and its data on reading scores in Subsections 1.2 and 1.3. One such question posed there was:

> *For British children aged 7 in 2004–2005, did boys' and girls' reading scores differ in location?*

Here, the two populations which we wish to compare in terms of their reading abilities are the population of British boys aged 7 in 2004–2005 and the population of British girls aged 7 in 2004–2005. Another example is the question

> *For British children aged 7–8 in 2004–2005, did reading scores differ in location according to their father's occupation?*

Here, the two populations which we wish to compare in terms of their reading abilities are the population of British children aged 7–8 in 2004–2005 whose father's occupation was coded 1 in Table 1 (managerial, technical, professional and skilled non-manual occupations) and the population of British children aged 7–8 in 2004–2005 whose father's occupation was coded 2 (skilled manual, partly skilled and unskilled occupations).

As in Section 5, comparisons will be made using hypothesis tests comparing means, and the two-sample z-test will be appropriate when both samples are large. To develop this test we use the reading scores from the BCS sample. We examine the first of the above questions:

> *For British children aged 7 in 2004–2005, did boys' and girls' reading scores differ in location?*

The following are appropriate null and alternative hypotheses:

> H_0: For British children aged 7 in 2004–2005, the mean reading score for girls is equal to the mean reading score for boys
>
> H_1: For British children aged 7 in 2004–2005, the mean reading score for girls is not equal to the mean reading score for boys.

We shall now introduce some symbols that will enable us to express our hypotheses more concisely and will also be helpful in explaining a theoretical result that we need. We are investigating two populations of values: the reading scores of all British 7-year-old girls in 2004–2005 and the reading scores of all British 7-year-old boys in 2004–2005. We shall let the means of these two populations be μ_g and μ_b, and the standard deviations be σ_g and σ_b. It is worth noting that the values of these quantities cannot be known: not all British 7-year-old girls and boys actually took this test in 2004–2005. So there is no way we could actually calculate μ_g, μ_b, σ_g and σ_b, but they enable us to make precise statements.

The subscripts 'g' and 'b' always relate to girls and boys, respectively.

For a start, we can use μ_g and μ_b to write the hypotheses concisely as

H_0: $\mu_g = \mu_b$
H_1: $\mu_g \neq \mu_b$,

or, equivalently, as

H_0: $\mu_g - \mu_b = 0$
H_1: $\mu_g - \mu_b \neq 0$.

This last form is the one we shall actually use to derive the test statistic.

Although we do not know test values for all children, the values for the samples of girls and boys in the BCS are known. We shall denote these samples' sizes by n_g and n_b, the sample means by $\overline{x}_g$ and $\overline{x}_b$, and the sample standard deviations by s_g and s_b. Their values were set out in Table 2 (Subsection 1.3), but we do not need them at the moment.

As we have expressed our null hypothesis as $\mu_g - \mu_b = 0$, it seems intuitively sensible to test the hypothesis by looking at the difference between the sample means, $\overline{x}_g - \overline{x}_b$. Before we can develop our hypothesis test, we need a theoretical result about the sampling distribution of the difference between two sample means.

You already know, from Section 4, that, because n_g and n_b are large, the sampling distribution of $\overline{x}_g$ is approximately normal with mean μ_g and standard error $\sigma_g/\sqrt{n_g}$, and similarly that the sampling distribution of $\overline{x}_b$ is approximately normal with mean μ_b and standard error $\sigma_b/\sqrt{n_b}$. We may conceive the first of these sampling distributions by thinking of all the possible samples of size n_g that we could select from the population of scores of all 7-year-old girls. We then imagine that we could calculate $\overline{x}_g$ for each of these samples and look at their distribution. Similar considerations apply to the sampling distribution of $\overline{x}_b$.

Now, think of all the possible means $\overline{x}_g$ of samples of size n_g of girls and also all the possible means $\overline{x}_b$ of samples of size n_b of boys. If we select just one value of $\overline{x}_g$ and one value of $\overline{x}_b$, we can calculate the difference $\overline{x}_g - \overline{x}_b$. Now think of all the possible pairs of values $\overline{x}_g$ and $\overline{x}_b$ we could select, and suppose we calculate $\overline{x}_g - \overline{x}_b$ for each of them. Then the distribution of all these differences is the **sampling distribution of the difference between two means**.

We require three results that are known about this sampling distribution. First, the mean of the sampling distribution of $\overline{x}_g - \overline{x}_b$ is equal to $\mu_g - \mu_b$, as you might expect. The second result requires the two samples to be independent of each other – here that is clearly the case, as the choice of girls was completely separate from the choice of boys. As long as the samples are independent, the standard deviation of the sampling distribution is given by

$$\text{SE} = \sqrt{\frac{\sigma_g^2}{n_g} + \frac{\sigma_b^2}{n_b}},$$

and this standard deviation is called the **standard error of the difference between two means**. Notice that it is larger than the

standard errors of $\overline{x}_g$ and $\overline{x}_b$, which are $\sigma_g/\sqrt{n_g}$ and $\sigma_b/\sqrt{n_b}$, respectively. This is because we are looking at the difference between two sample means; both means can vary, so there is more variation in the difference between them. Notice also that the standard error of the difference between two means is neither the sum nor the difference of the standard errors of the individual means. The next box summarises these results.

Mean and standard deviation of the sampling distribution of the difference between two means

- The mean of the sampling distribution is equal to $\mu_g - \mu_b$, the difference between the population means.
- The standard deviation of the sampling distribution is called the standard error of the difference between two means, and is given by

$$\text{SE} = \sqrt{\frac{\sigma_g^2}{n_g} + \frac{\sigma_b^2}{n_b}},$$

where n_g and n_b are the sizes of the samples, and σ_g and σ_b are the population standard deviations.

Furthermore, provided the sample sizes are sufficiently large, the sampling distribution of the differences between two sample means is approximately normal. This is the third result that we require.

Approximate normality of the sampling distribution of the difference between two means

If n_g and n_b are large, no matter what shape the population distributions, the sampling distribution of the difference between two means based on samples of sizes n_g and n_b will in practice be approximately normal.

From these results, $\overline{x}_g - \overline{x}_b$ is approximately normally distributed with mean $\mu = \mu_g - \mu_b$ and standard deviation $\sigma = \text{SE}$. Thus, the formula given in Subsection 3.3 can be used to transform $\overline{x}_g - \overline{x}_b$ to a quantity which follows (approximately) the standard normal distribution:

$$z = \frac{x - \mu}{\sigma} = \frac{(\overline{x}_g - \overline{x}_b) - (\mu_g - \mu_b)}{\text{SE}}.$$

Now to obtain our test statistic, we assume that the null hypothesis H_0 is true, so $\mu_g - \mu_b = 0$. We still cannot calculate z, as we do not know σ_g and σ_b. We deal with this problem exactly as we did in Subsection 5.1, by replacing σ_g by s_g and σ_b by s_b. This leads to the estimated standard error of $\overline{x}_g - \overline{x}_b$:

$$\text{ESE} = \sqrt{\frac{s_g^2}{n_g} + \frac{s_b^2}{n_b}}.$$

This logo suggests a more relaxing form of ESE

Test statistic and its sampling distribution when H_0 is true

For a two-sample z-test, when H_0: $\mu_g - \mu_b = 0$ is true, the test statistic,

$$z = \frac{\overline{x}_g - \overline{x}_b}{\text{ESE}}, \quad \text{where ESE} = \sqrt{\frac{s_g^2}{n_g} + \frac{s_b^2}{n_b}},$$

follows (approximately) the standard normal distribution.

For the one-sample z-test, we used the rule of thumb that the sample size had to be at least 25. To justify use of a two-sample z-test, we apply this rule of thumb to both samples and require that each sample size should be at least 25.

Since the test statistic above has the standard normal distribution (approximately) when the null hypothesis is true, the critical values are exactly the same as those in Subsection 5.1 for a one-sample hypothesis test. We can reject H_0 at the 1% significance level if $z \geq 2.58$ or if $z \leq -2.58$, and we can reject H_0 at the 5% significance level if $z \geq 1.96$ or $z \leq -1.96$. Otherwise we cannot reject H_0.

Example 8 *Comparing the mean reading scores of girls and boys*

We are now able to perform the two-sample z-test with which the current subsection was introduced. The hypotheses are:

H_0: $\mu_g = \mu_b$
H_1: $\mu_g \neq \mu_b$,

where μ_g is the population mean reading score for 7-year-old girls in 2004–2005, and μ_b is the population mean reading score for 7-year-old boys in 2004–2005. The data on which the test will be based were given as Table 2 (Subsection 1.3) and are repeated in Table 6.

Table 6 Summary statistics for data on reading scores of 7-year-old children

	Sample size	Sample mean	Sample standard deviation
Boys	206	109.31	27.671
Girls	190	113.42	25.464

(This data is copyright and owned by the Economic and Social Data Service.)

Key values for a two-sample z-test

In general, call the two groups A and B. The information you need to know for a two-sample z-test is:

- the sample means ($\overline{x}_A$ and $\overline{x}_B$)
- the sample sizes (n_A and n_B)
- the population standard deviations (σ_A and σ_B), or good estimates of them (s_A and s_B).

In this example we are using 'g' and 'b' to distinguish the two groups, rather than A and B. We have:

$$\overline{x}_{\mathrm{g}} = 113.42, \quad \overline{x}_{\mathrm{b}} = 109.31, \quad n_{\mathrm{g}} = 190, \quad n_{\mathrm{b}} = 206,$$
$$s_{\mathrm{g}} = 25.464, \quad s_{\mathrm{b}} = 27.671.$$

Both $n_{\mathrm{g}} = 190$ and $n_{\mathrm{b}} = 206$ are greater than 25, so we can assume that the z-test is applicable.

In the two-sample case, it is easier to calculate the value of z in two stages.

We first calculate the value of ESE, the estimated standard error of $\overline{x}_{\mathrm{g}} - \overline{x}_{\mathrm{b}}$:

$$\mathrm{ESE} = \sqrt{\frac{s_{\mathrm{g}}^2}{n_{\mathrm{g}}} + \frac{s_{\mathrm{b}}^2}{n_{\mathrm{b}}}} = \sqrt{\frac{25.464^2}{190} + \frac{27.671^2}{206}} \simeq 2.670.$$

Hence the value of the test statistic is

$$z = \frac{\overline{x}_{\mathrm{g}} - \overline{x}_{\mathrm{b}}}{\mathrm{ESE}} \simeq \frac{113.42 - 109.31}{2.670} \simeq 1.54.$$

The critical values are 1.96, -1.96 (5%) and 2.58, -2.58 (1%). Since $-1.96 < 1.54 < 1.96$, we cannot reject H_0 at the 5% significance level. There is little evidence to suggest that the mean reading scores in 2004–2005 for 7-year-old boys and girls were different.

The procedure for the two-sample z-test is summarised in the following box.

Procedure: two-sample z-test

1. Set up the null and alternative hypotheses,

   $$H_0\colon \mu_A = \mu_B$$
   $$H_1\colon \mu_A \neq \mu_B,$$

   where μ_A and μ_B are the means of populations A and B, respectively.

2. Calculate the test statistic

   $$z = \frac{\overline{x}_A - \overline{x}_B}{\text{ESE}},$$

   where the estimated standard error of $\overline{x}_A - \overline{x}_B$ is

   $$\text{ESE} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}.$$

   Here, n_A and n_B are the sample sizes of random samples from populations A and B respectively, $\overline{x}_A$ and $\overline{x}_B$ are the sample means, and s_A and s_B are the sample standard deviations.

3. Compare z with the appropriate critical values, which are 1.96 and -1.96 at the 5% significance level, and 2.58 and -2.58 at the 1% significance level.

   - If $z \geq 2.58$ or $z \leq -2.58$, then H_0 is rejected at the 1% significance level.
   - If $1.96 \leq z < 2.58$ or $-2.58 < z \leq -1.96$, then H_0 is rejected at the 5% significance level but not at the 1% significance level.
   - If $-1.96 < z < 1.96$, then H_0 is not rejected at the 5% significance level.

4. State the conclusions that can be drawn from the test.

In the two-sample z-test, it doesn't actually matter which of the two groups of interest you label A and which B. If you swapped the roles of A and B over, you would change the sign of z but nothing else. In particular, the conclusions of the test would be the same in either case.

Activity 27 *Mean reading scores of girls and boys at age 8*

In the BCS investigation, the following results were obtained for 8-year-old children.

Table 7 Summary statistics for data on reading scores of 8-year-old children

	Sample size	Sample mean	Sample standard deviation
Boys	145	126.38	29.927
Girls	138	127.49	25.064

(This data is copyright and owned by the Economic and Social Data Service.)

Carry out a two-sample z-test to investigate whether the mean reading score of 8-year-old girls in 2004–2005 was equal to the mean reading score of 8-year-old boys in 2004–2005. Comment on your result.

You might have noticed something interesting about the results for 7-year-old and 8-year-old children. For the younger children, the girls' sample mean score was $113.42 - 109.31 = 4.11$ more than that for boys, whereas for the older children the girls' sample mean score was $127.49 - 126.38 = 1.11$ higher. One might have thought at first glance that there was an interesting effect here: at the younger age, girls are ahead of boys in reading ability, but a year later boys seem to be catching up. Not so, however: our hypothesis tests showed that in neither case was there any evidence of a real difference, or therefore, of any such effect. The differences in the samples that we observed can easily have arisen by chance.

Does the level of education of parents have an affect on the reading scores of their children? In the next activity you will investigate this in the context of the BCS survey. This study classified parental education into two categories: those who finished full-time education by age 16 and those who continued after 16 (see Table 1, Subsection 1.2).

Activity 28 *Mean reading scores according to parental education*

This activity addresses another of the questions raised in Subsection 1.2:

> *For British children aged 7–8 in 2004–2005, did reading scores differ in location according to the level of their parents' education?*

Table 8 provides the relevant summary data from the BCS.

Table 8 Summary statistics for data on reading scores and parental education

Parental education	Sample size	Sample mean	Sample standard deviation
Ended by age 16	389	116.12	28.775
Continued after age 16	199	123.15	24.603

(This data is copyright and owned by the Economic and Social Data Service.)

(Note that 679 children were tested for their reading ability, but no information was available on when the parents of 91 of the children completed their education.)

Carry out a hypothesis test to investigate whether children whose parents' education continued beyond age 16 scored differently on average on the reading test from those children whose parents' education ended by age 16.

You might have expected the answer to Activity 28 before analysing the data. That is, denoting μ_C as the mean reading score of children whose parental education continued after age 16 and μ_E as the mean reading score of children whose parental education ended by age 16, you might have thought of doing the following: testing the null hypothesis that $\mu_C = \mu_E$ with the purpose of seeing whether, as you suspect, μ_C is actually greater than μ_E, disregarding the possibility that μ_C could be less than μ_E. Hypothesis tests undertaken when a particular type of inequality between the two groups is of interest are the one-sided tests mentioned in a margin note at the start of Section 5 and to be looked at briefly in Unit 10.

You have now covered the material related to Screencast 6 for Unit 7 (see the M140 website).

Exercises on Section 6

Exercise 19 *Mean reading scores according to fathers' occupations*

This exercise concerns another question posed in Subsection 1.2, namely:

> *For British children aged 7–8 in 2004–2005, did reading scores differ in location according to their fathers' occupations?*

Table 9 provides the relevant summary data from the BCS. Note that, as in Table 1 (Subsection 1.2), '1' denotes 'managerial, technical, professional and skilled non-manual' occupations while '2' denotes 'skilled manual, partly skilled and unskilled' occupations.

Table 9 Summary statistics for data on reading scores and father's occupation

Father's occupation	Sample size	Sample mean	Sample standard deviation
1	316	120.55	24.221
2	203	117.17	30.085

(This data is copyright and owned by the Economic and Social Data Service.)

(No information was available on father's occupation for 160 individuals.)

Carry out a two-sample hypothesis test to investigate whether the mean score of children whose father had an occupation coded 1 differs from that of children whose father had an occupation coded 2. Comment on your result.

Exercise 20 *Calcium for babies*

This exercise is related to an investigation of the effect of vitamin D supplementation for the prevention of low levels of calcium in newborn babies. The data given in Table 10 come from a clinical trial in which a sample of babies who were breast-fed were compared with a sample of babies who were bottle-fed: the measured quantity was the level of calcium in the baby's blood ('serum calcium') at 1 week of age.

Table 10 Summary statistics for data on serum calcium for week-old babies

	Sample size	Sample mean	Sample standard deviation
Breast-fed	64	2.45	0.292
Bottle-fed	169	2.30	0.274

(Source: Cockburn et al. (1980) 'Maternal vitamin D intake and mineral metabolism in mothers and their newborn infants', *British Medical Journal*, vol. 281, pp. 11–14)

Carry out a two-sample z-test to investigate whether the mean serum calcium level of week-old babies was the same whether they were breast-fed or bottle-fed.

Exercise 21 *Peak flow rate of lungs*

The peak flow rate is a measure of how well a person's lungs are functioning. It is the maximum rate in litres per minute at which air can be expelled through a peak flow meter. In an investigation of the possibility that chronic bronchitis, although a disease of adult life, starts in childhood, the peak flow rates of a large number of school children without persistent coughs were measured. Amongst other details recorded were whether the child lived in an urban or a rural area. Data for urban and rural areas are summarised in Table 11. Use a two-sample z-test to examine whether the average peak flow rate of children differs in these two groups.

Table 11 Peak flow rates for children without persistent coughs

	Sample size	Sample mean	Sample standard deviation
Urban	485	226	52
Rural	637	231	53

(Source: unpublished data collected by Professor J.R.T. Colley, University of Bristol)

7 Computer work: one-sample z-tests

In this section you will use Minitab to perform one-sample z-tests. These are similar to the tests you have performed earlier in this unit, except that Minitab gives the results of hypothesis tests in terms of p-values, while in earlier sections we have only considered specific significance levels (5% and 1% significance levels). The use of p-values with sign tests was explained in Unit 6. Their use with z-tests is identical, but is described explicitly in the Computer Book.

You should now turn to the Computer Book and work through Chapter 7. The chapter starts with the interactive computer resources connected with Section 3 of this unit; you should do them now if you have not already done so. You should then do the Minitab work that is contained in the rest of Chapter 7.

8 Conclusions and reservations

We have answered many of the questions raised in Section 1, and we have learned a lot about children's reading ability and factors affecting it, at any rate for British children aged 7 and 8 in 2004–2005. We summarise our conclusions below. As usual, though, after coming to such conclusions, we should stop and look for reservations that might arise.

- Are there any problems with the data that might throw doubt on conclusions drawn from them?
- Were appropriate statistical methods used in analysing the data?

We shall look at both these questions. To address the second question we shall discuss when z-tests should be used. We then note limitations on the way conclusions are stated and interpreted.

Conclusions

We began this unit by asking the general question:

What factors affect a child's reading ability?

In Section 1, we refined this question to produce several more specific questions that we could attempt to answer using BCS data. In Sections 5 and 6, we carried out hypothesis tests that related to these questions. All these tests involved hypotheses about the population from which the BCS sample was drawn, that of British children aged 7 and 8 in 2004–2005.

In Example 7 (Subsection 5.2), we found that we could reject the null hypothesis that 7-year-old British children in 2004–2005 had the overall population mean reading score for 7-year-olds. Similarly, in Activity 25 (Subsection 5.2), we found that we could reject the null hypothesis that 8-year-old British children in 2004–2005 had the overall population mean reading score for 8-year-olds. (A related result for 7-year-old girls was obtained in Exercise 16 in Section 5.)

In Example 8 (Section 6), we found that, for 7-year-old children, the null hypothesis that the population mean for boys was equal to that for girls could not be rejected. In Activity 27 (Section 6), we also found that the same was true for 8-year-old children.

In Activity 28 (Section 6), we found strong evidence that the mean reading score was higher for children whose parents' education had lasted longer. (Something less expected happened with respect to father's occupation in Exercise 19.)

Reservations about the data

Probably the main reservation about the data is whether they can be considered a random sample from the relevant population. As was discussed in Activity 3 (Subsection 1.2), the data do not come from a formal random sample of British children aged 7 and 8 in 2004–2005, of the sort one might draw using a sampling frame and random numbers. But it might still be the case that the data can be *treated* as if they had been drawn in that way. How would a 'real' random sample differ from the BCS 2004–2005 sample? The main difference was raised in Activity 3; all the children in our sample have at least one parent who is in the BCS survey, and therefore was born in a particular week in 1970. In a true random sample of British 7- and 8-year-old children, not every child would have a parent aged 34.

There are other features of the sampling process that might lead to the sample of children being unrepresentative:

- Children could be included only if the BCS 2004–2005 investigators had managed to trace their parents. People in the original BCS sample whose lifestyles involve moving around a lot may have been harder to trace, and therefore their children would be less likely to be in the sample.
- There are missing data. This data may not be missing completely 'at random' – which might be OK, provided there is not too much of it – but its very missingness might be connected to the things you are trying to measure. (This is a common problem in real-world statistics.)

  For example, parents with less education *might* be more reluctant to say so in response to a survey, in which case children with such parents might be under-represented; worse, such parents might be more likely to not respond to the education question if they know their child is not reading especially well, and they don't want to be 'blamed' for this situation.
- The data on parental education simply give the age at which *one* of the parents left full-time education, and say nothing about which parent it was. Also, nothing is said about any qualifications he or she gained, or about any part-time study.

These reservations about the randomness/representativeness of the sample are probably less important than the reservation about the parents' age, but they should not be forgotten.

Any other reservations?

8.1 When to use the z-test

The z-test can be applied in many situations, though it does have limitations. In this subsection the characteristics of the test are described so that you can recognise when it is appropriate.

The sample size must be large

It is unnecessary to know anything about the distribution of the population from which the sample is selected, because the test is based on the fact that the sampling distribution of the mean of a sample of size n is approximately normal, provided n is sufficiently large.

As a simple rule of thumb, we assume that in the one-sample case, n should be at least 25, and in the two-sample case, n_A and n_B should both be at least 25. If the sample size is less than 25, you should not apply the z-test. (If you believe that the population distribution is extremely skew – which has not been the case for any distribution in this unit – then it is safer only to use the z-test if the sample size is considerably greater than 25.)

In Unit 10 you will meet another hypothesis test, the t-test, which you can apply under some circumstances when the sample size is less than 25.

The sample values should consist of numerical measurements

The z-test should be applied only to data which consist of numerical measurements. Length, weight, time, scores in a test and petrol consumption are all examples of such data. The z-test cannot be applied, for example, to data which might be coded, such as perhaps hair colour or disease type – with such data the concept of a population mean or a sample mean is not really meaningful.

The samples should be unrelated

This restriction applies only to the two-sample z-test. The samples from the two populations should be unrelated and so not consist of data collected in pairs, each pair coming from the same individual.

All the hypothesis tests that we have performed in this unit were based on data that met the requirements for the one- or two-sample z-tests. Sample sizes were above 25 (substantially so for the BCS data), sample values were numerical measurements (often scores on a reading test), and the two-sample z-test was only ever applied to unrelated samples from two separate populations.

8.2 Limitations in stating conclusions

In stating conclusions from any hypothesis test, the following factors must be borne in mind.

- A sampling error may have occurred.
- The conclusions should match the population from which the sample was drawn.
- The conclusions must not make causal statements which are not supported by the way the data arose.

Let us look at each of these briefly.

Sampling errors

You should always bear in mind that a sampling error might have occurred; that is, the result of any hypothesis test might be due to sampling variation. Hypothesis tests do not provide *proofs* of the truth of either the null or alternative hypotheses. They just attempt to assess the evidence for or against the hypotheses. For example, if the null hypothesis is rejected, that means that there is evidence against the null hypothesis, but not that the null hypothesis is definitely wrong. However, with the BCS data, in most of the hypothesis tests where we rejected the null hypothesis, the test statistic came out much higher numerically than the critical values (it easily gave 'strong evidence'), so with those tests it is unlikely – but still *possible* – that sampling error has led to erroneous conclusions.

What can we say about the populations?

A major difficulty with the BCS data is that it is not clear that these data can be treated as a random sample from *any* population. But they are clearly likely to be much more representative of the population of British children than of, say, Ugandan children. The stated conclusions were explicit in referring to British children and to the year, 2004–2005, in which the data were collected, although we should perhaps have referred to the population of British children, aged 7 or 8 in 2004–2005, *who had at least one parent aged 34*, as all these characteristics are common to the children in our sample. Nevertheless, it seems reasonably plausible that the data would still be representative of British children in some other year close to 2004–2005, say 2003 or 2007, since reading skills are unlikely to change very rapidly. But it would be a mistake to apply the conclusions directly to the population of British children in 1970, say, or 2013.

What can we say about causal statements?

Causality will be discussed at much greater length in another hypothesis testing context in Unit 8.

Can we make any conclusions about what might have *caused* any differences for which we have evidence? For the BCS data, the answer is, essentially, 'no'! Our conclusions are not worded in causal terms; for instance, we concluded (in Activity 28) that, for British children aged 7 and 8 in 2004–2005, those whose parental education was beyond the age of 16 had a higher mean reading score than did those whose parent left education earlier. Worded like that, the conclusion says nothing about how this difference arose; but there is a great temptation to suppose that the parent's level of education *caused* the difference in mean reading score.

This causal conclusion goes beyond what the data tell us. Instead, there could well be one or more other factors that underly both a child's reading ability and whether a parent of the child was educated past the age of 16. We just cannot tell about such things from these data, since they do not give us the appropriate information.

Summary

In terms of statistical methodology, you have been introduced to the most important distribution in statistics – the normal distribution – and you have learned to use the distribution in two hypothesis tests, the one-sample and two-sample z-tests. In this unit, the normal distribution arose out of consideration of the sampling distributions of the sample mean: regardless of the distribution of the original data, such sampling distributions were seen to become more and more normal-like as the sample size, n, increased. You then learned about the normal distribution itself. You saw the way in which it depends on two quantities, the population mean, μ – controlling its location – and the population standard deviation, σ – controlling its spread. You also learned how any normal distribution can be related to a special normal distribution: the standard normal distribution with $\mu = 0$ and $\sigma = 1$. You then found that the sampling distribution of the sample mean can be approximated by a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$, which is called the standard error of the mean.

The z-test was first introduced in its one-sample form to address null and alternative hypotheses concerning the value of μ. Its test statistic was developed in two forms, for σ assumed to be known and, more usefully, for σ unknown. You saw how the sampling distribution of the test statistic, and hence the critical values associated with the test, arose from the above results for the normal distribution. Having learned how to implement the one-sample z-test, you went on to learn how to adapt those ideas to produce the two-sample z-test; this is applicable to testing hypotheses concerning whether or not the means of two unrelated populations are equal. In each case, you applied what you learned about hypothesis testing in Unit 6 to interpret results in terms of the amount of evidence the data provide against the null hypothesis.

What you learned about children's reading abilities from the BCS survey has been summarised and discussed in Section 8.

Learning outcomes

After you have worked through this unit, you should be able to:

- appreciate the steps taken to make the unit's original question, which is rather general, more specific
- recall that the null and alternative hypotheses required for the z-test are expressed in terms of population means
- recognise a bell-shaped distribution
- appreciate that population distributions can have different shapes, some of which are normal
- appreciate that, whatever the shape of the population distribution, for a large enough sample size the sampling distribution of the mean is nearly always approximately normal
- appreciate the relationship between the location and spread of a normal distribution and its mean and standard deviation
- appreciate that it makes sense to think of normal distributions in terms of the number of standard deviations of the variable away from its mean, and that we can therefore think of all normal distributions in terms of only one distribution: the standard normal distribution
- apply the formula that transforms any variable x with a given normal distribution to the variable z with the standard normal distribution
- understand what is meant by the standard error (of the mean) and the estimated standard error in one- and two-sample situations
- write down the mean and standard deviation of the sampling distribution of the mean for samples of size n, given the population mean, μ, and standard deviation, σ
- follow the reasoning behind the one-sample z-test and apply the test when σ is assumed known
- adapt and apply the one-sample z-test when σ is unknown
- understand and apply the two-sample z-test to analyse the difference between means
- use Minitab to perform the one-sample z-test
- be aware of questions to ask which might lead to reservations about the conclusions of a hypothesis test
- be aware of some of the characteristics of the z-test, and recognise when it is necessary to exercise some caution in its use.

Solutions to activities

Solution to Activity 1

There are many possible answers to this question. You can test a child's reading ability by how well they read a coherent passage, recognise separate words, name letters, or pronounce separate words. Perhaps you have thought of other measures; or you may have thought in terms of a standard reading test of some kind.

Solution to Activity 2

Some of the factors you may have thought of are pre-school education, parents' education, precise age of child, whether there are other children in the family, mental or physical disability, social deprivation, quality of teaching, method of teaching, school class size, and parent's reading to the child at an early age. You may have been able to think of a different set of possibilities.

Solution to Activity 3

To be a random sample of exactly the sort you met in Unit 4, the sample would have had to be chosen by using random numbers to select children from a sampling frame of all 7- and 8-year-old children in the country. Clearly this was not done, so in this sense the sample is not random. However, you have previously met examples where a sample that was *not* chosen in this way was nevertheless considered to be representative in the same way that a formally selected random sample would be. In other words, the key question is not 'Was this sample chosen using a sampling frame and random numbers?', but 'Was this sample chosen in such a way that it has the same properties as one chosen using a sampling frame and random numbers?'

The answer to the second question is not so clear in this case. It might seem reasonable to treat the original BCS sample of people born in a particular week in 1970 as being representative of the general population of people born in Great Britain around that time, in the same way that a random sample would be representative. It is perhaps less reasonable to treat their 7- and 8-year-old children as if they were a random sample from the population of all 7- and 8-year-old children in 2004–2005. This is because in a true random sample of children, the ages of the children's parents would vary more – in this sample all the children have at least one parent born in a particular week in 1970. This might be quite a problem because the age and experience of their parents might well be linked to how a child's reading develops.

Solution to Activity 4

(a) The 7-year-old boys are identified in Table 1 by having a value of 1 in the third column (Gender – 1 denotes boy) and 1 in the fourth column (Coded age – 1 denotes 7 years old). There are six individuals in Table 1 that have 1 in each of the third and fourth columns. They have reading scores

106 110 134 25 172 160.

The sample size is $n = 6$.

(b) To calculate $\overline{x}$,

$$\sum x = 106 + 110 + 134 + 25 + 172 + 160 = 707,$$

and so

$$\overline{x} = \frac{\sum x}{n} = \frac{707}{6} \simeq 117.8.$$

Using Method 2 from Unit 3 (Subsection 3.1) to calculate s,

$$\begin{aligned}\sum (x - \overline{x})^2 &= \sum x^2 - \frac{(\sum x)^2}{n} \\ &= 97\,101 - \frac{(707)^2}{6} \\ &\simeq 13\,792.833.\end{aligned}$$

This means that the variance is

$$\begin{aligned}\frac{\sum (x - \overline{x})^2}{n - 1} &= \frac{13\,792.833}{5} \\ &\simeq 2758.5667.\end{aligned}$$

So

$$\begin{aligned}s = \sqrt{\text{variance}} &= \sqrt{2758.5667} \\ &\simeq 52.5.\end{aligned}$$

Solution to Activity 5

The proportion of students on MS221 in the presentation in question achieving 75 marks is the actual number of students receiving 75 marks (21) divided by the total number of students sitting the exam (1234). That is,

$$\frac{21}{1234} \simeq 0.0170.$$

Solution to Activity 6

(a) (i) Sample mean $= \dfrac{15+35}{2} = \dfrac{50}{2} = 25.$

(ii) Sample mean $= \dfrac{65+77}{2} = \dfrac{142}{2} = 71.$

(iii) Sample mean $= \dfrac{65+52}{2} = \dfrac{117}{2} = 58.5.$

(iv) Sample mean $= \dfrac{37+80}{2} = \dfrac{117}{2} = 58.5.$

(b) The sample means of samples of size 2 are either integers (as in (i) and (ii) in part (a)) or else 'half-integers', that is, values of the form 'an integer plus a half' (as in (iii) and (iv) of part (a)).

Solution to Activity 7

The distribution of sample means of size 2 shown in Figure 3 is much smoother and less jagged than the distribution of the population data shown in Figure 2. The distribution of sample means of size 2 is also fairly symmetric, about a maximal value at around 70. However, there are slightly more sample means less than 70 than greater than 70, meaning that the distribution is slightly left-skew (see Subsection 5.2 of Unit 1). You might also note that the distribution fades away to almost nothing – corresponding to very rare sample mean values – at about 10 or so.

Solution to Activity 8

(a) Sample mean $= \dfrac{10+20+45}{3} = \dfrac{75}{3} = 25.$

(b) Sample mean $= \dfrac{82+24+33}{3} = \dfrac{139}{3} \simeq 46.3.$

(c) Sample mean $= \dfrac{52+61+73}{3} = \dfrac{186}{3} = 62.$

(d) Sample mean $= \dfrac{78+64+46}{3} = \dfrac{188}{3} \simeq 62.7.$

Solution to Activity 9

The distribution of sample means of size 3 shown in Figure 4 is much smoother than the distribution of sample means of size 2 shown in Figure 3 – it is made up of many more very short lines whose overall effect is closer to a smooth curve. The sampling distribution in Figure 4 is a little more compressed from side to side than that in Figure 3; that is, it has a smaller spread. The sampling distribution in Figure 4 is perhaps even closer to symmetric than the one in Figure 3. The maximum value about which the sampling distribution is approximately symmetric is, however, at approximately the same place as the maximum in Figure 3 – that is, at about, or a little under, 70. Finally, corresponding to its smaller spread, the distribution in Figure 4 fades away to almost nothing at about 20 or so (and just below 100).

Solution to Activity 10

The spread of the sampling distribution in Figure 5 is a little smaller again than the spread of the sampling distribution in Figure 4. It is also the case that any skewness apparent in Figure 4 is no longer apparent in Figure 5: this time, the distribution is symmetric, falling away smoothly on either side of a maximum value a little way below 70. But aside from the change in spread, the sampling distribution in Figure 5 is rather similar to the sampling distribution in Figure 4; in particular, the maximum is at approximately the same place in the two figures, while in both cases the sampling distributions fall away from the maximum, first more rapidly and then more slowly as they 'level out' a long way from the maximum.

Solution to Activity 11

As the sample size n increases, the sampling distributions, which all have the same symmetric shape, rise more and more sharply to a mode (at a little below 70, it seems). Also, the distributions become more and more compressed (i.e. the spread decreases as the sample size increases).

Solution to Activity 12

For $n = 2$, the sampling distribution of the mean is right-skew, but a little less so than the population distribution. As the sample size n increases, the sampling distributions again become more symmetric and bell-shaped. The distributions also become more and more peaked and compressed about the mode (at about £500).

Solution to Activity 13

The centre of this normal distribution is located at the value 1, so, as in Figure 11(b), this means that $\mu = 1$. The distribution also appears to have the same spread as the normal distribution in Figure 12(c), so $\sigma = 2$. To confirm these claims, notice that the x-axis labels on Figure 11(b) have 1 added to them (when $\mu = 1$) compared with the corresponding labels on Figure 11(a) (when $\mu = 0$); similarly, the x-axis labels on Figure 13 have 1 added to them (when $\mu = 1$) compared with the corresponding labels on Figure 12(c) (when $\mu = 0$).

Don't worry if you didn't get this activity right. There is much more on changing both μ and σ in the normal distribution in the Computer Book and Subsections 3.2 and 3.3 to follow.

Solution to Activity 14

(a) The mode of this normal distribution occurs at about $x = 10$. So $\mu \simeq 10$. Almost all the distribution is contained between $x = 4$ and $x = 16$ (i.e. within 10 ± 6). So $3\sigma \simeq 6$ and $\sigma \simeq 2$. That is, the normal distribution plotted in Figure 17 is approximately the normal distribution with mean $\mu = 10$ and standard deviation $\sigma = 2$.

(b) The mode of this normal distribution occurs at about $x = 100$. So $\mu \simeq 100$. Almost all the distribution is contained between $x = 40$ and $x = 160$ (i.e. within 100 ± 60). So $3\sigma \simeq 60$ and $\sigma \simeq 20$.

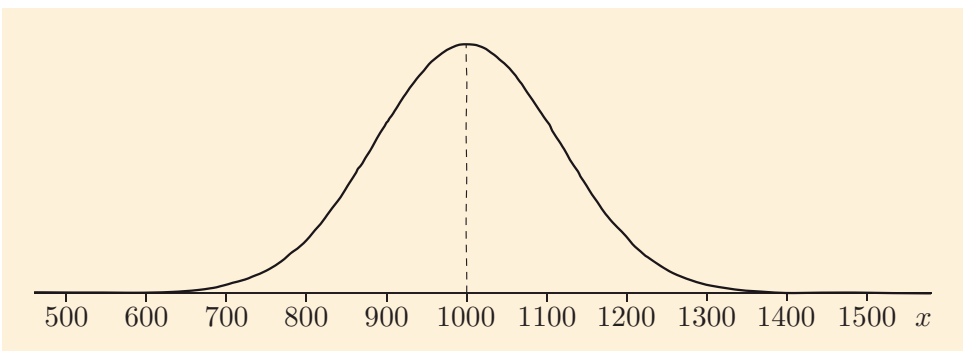
That is, the normal distribution plotted in Figure 18 is approximately the normal distribution with mean $\mu = 100$ and standard deviation $\sigma = 20$.

(c) The mode of this normal distribution occurs at about $x = 1$. So $\mu \simeq 1$. Almost all the distribution is contained between $x = 0.7$ and $x = 1.3$ (i.e. within 1 ± 0.3). So $3\sigma \simeq 0.3$ and $\sigma \simeq 0.1$. That is, the normal distribution plotted in Figure 19 is approximately the normal distribution with mean $\mu = 1$ and standard deviation $\sigma = 0.1$.

Solution to Activity 15

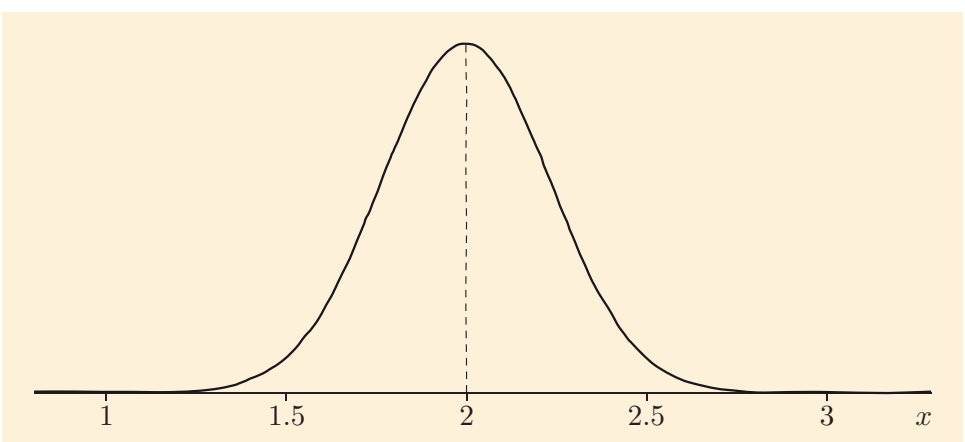
You should have obtained something like the sketches below, although, since you may have used different scales, yours could look a bit different. The important thing is that the information on your horizontal axes should match those in the figures.

- The following normal distribution is centred at $\mu = 1000$ and has just about all the distribution contained within $1000 \pm (3 \times 100) = 1000 \pm 300$, i.e. between 700 and 1300.



The normal distribution with $\mu = 1000$, $\sigma = 100$

- The following normal distribution is centred at $\mu = 2$ and has almost all the distribution contained within $2 \pm (3 \times 0.25) = 2 \pm 0.75$, i.e. between 1.25 and 2.75.



The normal distribution with $\mu = 2$, $\sigma = 0.25$

Solution to Activity 16

(a) Here $\mu = 10$ and $\sigma = 2$, so

$$z = \frac{x - 10}{2}.$$

(b) Here $\mu = 100$ and $\sigma = 20$, so

$$z = \frac{x - 100}{20}.$$

(c) Here $\mu = 1$ and $\sigma = 0.1$, so

$$z = \frac{x - 1}{0.1}.$$

If you prefer, you could equivalently write this as

$$z = \frac{x - 1}{1/10} = 10(x - 1).$$

Solution to Activity 17

(a) The appropriate formula is

$$z = \frac{h - \mu}{\sigma},$$

where $\mu = 1.75$ and $\sigma = 0.07$. Hence

$$z = \frac{h - 1.75}{0.07}.$$

(b) When $h = 1.96$,

$$z = \frac{1.96 - 1.75}{0.07} = \frac{0.21}{0.07} = 3.$$

So a height of 1.96 metres is 3 standard deviations above the mean height of 1.75 metres.

When $h = 1.61$,

$$z = \frac{1.61 - 1.75}{0.07} = \frac{-0.14}{0.07} = -2.$$

So a height of 1.61 metres is 2 standard deviations below the mean height of 1.75 metres.

When $h = 1.785$,

$$z = \frac{1.785 - 1.75}{0.07} = \frac{0.035}{0.07} = 0.5.$$

So a height of 1.785 metres is 0.5 standard deviations above the mean height of 1.75 metres.

You can check the picture of the distribution in Figure 14 (Subsection 3.2) to see if each of the values of h in this activity is the appropriate z standard deviations away from the mean.

Solution to Activity 18

(a) In each case the sampling distribution is symmetric about a mode at about 66 marks. So the means of the sampling distributions appear to be the same as the population mean $\mu = 66$ marks.

(b) The sampling distributions all look symmetric with a mode at about £491 or so. So again the mean of each of the sampling distributions appears to be the same as the population mean $\mu =$ £491.

Solution to Activity 19

(a) The standard deviation of the sampling distribution of mean exam marks decreases (i.e. the distributions become more compressed) as the sample size n increases.

(b) The standard deviation of the sampling distribution of mean employees' earnings also decreases (i.e. the distributions become more compressed) as the sample size n increases.

Solution to Activity 20

(a) When $n = 25$,

$$\frac{\sigma}{\sqrt{n}} = \frac{22}{\sqrt{25}} = 4.4,$$

(b) When $n = 50$,

$$\frac{\sigma}{\sqrt{n}} = \frac{22}{\sqrt{50}} \simeq 3.11.$$

(c) When $n = 100$,

$$\frac{\sigma}{\sqrt{n}} = \frac{22}{\sqrt{100}} = 2.2.$$

Solution to Activity 21

When $n = 25$ and $\sigma = 0.01$,

$$\frac{\sigma}{\sqrt{n}} = \frac{0.01}{\sqrt{25}} = \frac{0.01}{5} = 0.002.$$

It follows that the sampling distribution of the mean for samples of 25 ball bearings from this manufacturer is approximately normal with mean $\mu = 2\,\text{mm}$ and standard deviation $\sigma/\sqrt{n} = 0.002\,\text{mm}$.

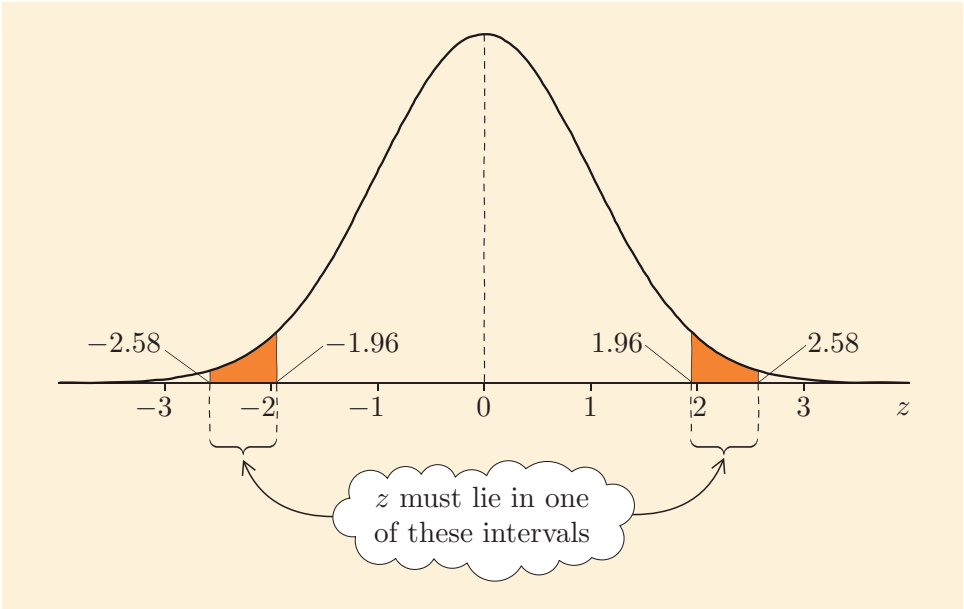
Solution to Activity 22

The value of z is

$$z = \frac{\overline{x} - A}{\text{SE}} = \frac{112 - 120}{15/\sqrt{100}} \simeq -5.33.$$

Solution to Activity 23

If H_0 is rejected at the 5% significance level but not at the 1% significance level, then z lies in the critical region shown in Figure 32 but not in the critical region shown in Figure 33, that is $1.96 \leq z < 2.58$ or $-2.58 < z \leq -1.96$. This is shown in the following figure.



Sketch of standard normal distribution with possible values of z indicated

Solution to Activity 24

(a) The null and alternative hypotheses are

$$H_0\colon\ \mu = 26.1$$
$$H_1\colon\ \mu \neq 26.1,$$

where μ is the population mean set-up time of the new method.

(b) $A = 26.1$, as this is the value of μ under H_0. The sample values are $\overline{x} = 20.9$ and $n = 53$.

(c) The test statistic is

$$z = \frac{\overline{x} - A}{\text{SE}} = \frac{20.9 - 26.1}{12.3/\sqrt{53}} \simeq -3.08.$$

(d) As -3.08 is less than -1.96 and -2.58, the null hypothesis is rejected at both the 5% significance level and the 1% significance level.

(e) There is strong evidence against H_0. Thus there is strong evidence that the mean set-up time under the new method differs from that under the old method – there is strong evidence that the new method is faster.

Solution to Activity 25

The appropriate null and alternative hypotheses are

$$H_0\colon\ \mu = 116$$
$$H_1\colon\ \mu \neq 116,$$

where μ is the population mean reading score of all British 8-year-old children in 2004–2005.

As the sample size, 283, is much greater than 25, it is appropriate to apply the z-test. We have

$$A = 116, \quad \overline{x} = 126.92, \quad n = 283, \quad s = 27.711.$$

The test statistic is

$$z = \frac{\overline{x} - A}{\text{ESE}} = \frac{\overline{x} - A}{s/\sqrt{n}} = \frac{126.92 - 116}{27.711/\sqrt{283}} \simeq 6.63.$$

The critical values are 1.96, -1.96 (5%) and 2.58, -2.58 (1%). Since $6.63 \geq 2.58$, we can reject the null hypothesis at the 1% significance level and conclude that there is strong evidence that the mean reading score of 8-year-olds in 2004–2005 was not equal to 116.

This might have been a little surprising if we had not seen a similar result for 7-year-old children in Example 7.

Solution to Activity 26

The null and alternative hypotheses are

$$H_0\colon\ \mu = 381.50$$
$$H_1\colon\ \mu \neq 381.50,$$

where μ is the population mean weekly wage (in £) of female local government clerical officers and assistants in 2011.

As the sample size, 810, is greater than 25, it is appropriate to apply the z-test. We have

$$A = 381.5, \quad \overline{x} = 373.4, \quad n = 810, \quad s = 138.2.$$

The test statistic is

$$z = \frac{\overline{x} - A}{\text{ESE}} = \frac{\overline{x} - A}{s/\sqrt{n}} = \frac{373.4 - 381.5}{138.2/\sqrt{810}} \simeq -1.67.$$

The critical values are 1.96, -1.96 (5%) and 2.58, -2.58 (1%). Since $-1.96 < -1.67 < 1.96$, the null hypothesis is not rejected at the 5% significance level. There is little evidence that the mean weekly wage of female local government clerical officers and assistants differed from the overall mean weekly wage of female employees in 2011.

Solution to Activity 27

The null and alternative hypotheses are

$$H_0\colon\ \mu_{\text{g}} = \mu_{\text{b}}$$
$$H_1\colon\ \mu_{\text{g}} \neq \mu_{\text{b}},$$

where μ_{g} and μ_{b} are the population mean reading scores for 8-year-old girls and boys, respectively. We have:

$$\overline{x}_{\text{g}} = 127.49, \quad \overline{x}_{\text{b}} = 126.38, \quad n_{\text{g}} = 138, \quad n_{\text{b}} = 145,$$
$$s_{\text{g}} = 25.064, \quad s_{\text{b}} = 29.927.$$

Both $n_{\text{g}} = 138$ and $n_{\text{b}} = 145$ are greater than 25, so we can assume that the z-test is applicable.

The estimated standard error is

$$\text{ESE} = \sqrt{\frac{s_{\text{g}}^2}{n_{\text{g}}} + \frac{s_{\text{b}}^2}{n_{\text{b}}}} = \sqrt{\frac{25.064^2}{138} + \frac{29.927^2}{145}} \simeq 3.276,$$

and the test statistic is

$$z = \frac{\overline{x}_{\text{g}} - \overline{x}_{\text{b}}}{\text{ESE}} \simeq \frac{127.49 - 126.38}{3.276} \simeq 0.34.$$

The critical values are 1.96, −1.96 (5%) and 2.58, −2.58 (1%). Since $-1.96 < 0.34 < 1.96$, we cannot reject the null hypothesis at the 5% significance level. There is no reason to doubt that the mean reading scores of 8-year-old girls and boys were the same.

Solution to Activity 28

Let 'E' denote quantities relating to children whose parental education ended by age 16, and 'C' denote quantities relating to children whose parental education continued after age 16. The null and alternative hypotheses are

$$H_0\colon\ \mu_{\text{C}} = \mu_{\text{E}}$$
$$H_1\colon\ \mu_{\text{C}} \neq \mu_{\text{E}},$$

where μ_{C} and μ_{E} are the population mean reading scores of interest. We have:

$$\overline{x}_{\text{E}} = 116.12, \quad \overline{x}_{\text{C}} = 123.15, \quad n_{\text{E}} = 389, \quad n_{\text{C}} = 199,$$
$$s_{\text{E}} = 28.775, \quad s_{\text{C}} = 24.603.$$

Both $n_{\text{C}} = 199$ and $n_{\text{E}} = 389$ are greater than 25, so we can assume that the z-test is applicable.

The estimated standard error is

$$\text{ESE} = \sqrt{\frac{s_{\text{C}}^2}{n_{\text{C}}} + \frac{s_{\text{E}}^2}{n_{\text{E}}}} = \sqrt{\frac{24.603^2}{199} + \frac{28.775^2}{389}} \simeq 2.274,$$

and the test statistic is

$$z = \frac{\overline{x}_{\text{C}} - \overline{x}_{\text{E}}}{\text{ESE}} \simeq \frac{123.15 - 116.12}{2.274} \simeq 3.09.$$

The critical values are 1.96, −1.96 (5%) and 2.58, −2.58 (1%). Since $3.09 \geq 2.58$, we can reject H_0 at the 1% significance level. There is strong evidence that the mean reading score of children whose parental education continued after age 16 differs from the mean reading score of children whose parental education ended by age 16. There is strong evidence that the children of parents who stayed longer in full-time education did better than those of parents who left education earlier.

Solutions to exercises

Solution to Exercise 1

(a) The 8-year-old children are identified in Table 1 by having a value of 2 in the fourth column (Coded age '2' denotes 8 years old). There are four individuals in Table 1 that have 2 in the fourth column. They have reading scores

118 115 56 136.

The sample size is $n = 4$.

(b) To calculate $\overline{x}$,

$$\sum x = 118 + 115 + 56 + 136 = 425,$$

and so

$$\overline{x} = \frac{\sum x}{n} = \frac{425}{4} = 106.25.$$

To calculate s,

$$\begin{aligned}\sum (x - \overline{x})^2 &= \sum x^2 - \frac{(\sum x)^2}{n} \\ &= 48\,781 - \frac{425^2}{4} \\ &= 3624.75,\end{aligned}$$

which means the variance is

$$\frac{\sum (x - \overline{x})^2}{n - 1} = \frac{3624.75}{3} = 1208.25.$$

So

$$s = \sqrt{\text{variance}} = \sqrt{1208.25} \simeq 34.8.$$

Solution to Exercise 2

(a) The children of interest in this exercise are identified in Table 1 by having a value of 1 in the fifth column (Parental education '1' denotes finished aged 16 or less) and a value of 1 in the sixth column (Father's occupation '1' denotes managerial, technical, professional and skilled non-manual occupations). There are seven individuals in Table 1 that have 1 in both the fifth and sixth columns. They have reading scores

123 110 134 110 172 136 160.

The sample size is $n = 7$.

(b) To calculate $\overline{x}$,

$$\sum x = 123 + 110 + 134 + 110 + 172 + 136 + 160 = 945,$$

and so

$$\overline{x} = \frac{\sum x}{n} = \frac{945}{7} = 135.$$

To calculate s,

$$\sum(x-\overline{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$
$$= 130\,965 - \frac{945^2}{7}$$
$$= 3390,$$

which means the variance is

$$\frac{\sum(x-\overline{x})^2}{n-1} = \frac{3390}{6} = 565.$$

So

$$s = \sqrt{\text{variance}} = \sqrt{565} \simeq 23.8.$$

Solution to Exercise 3

Suitable null and alternative hypotheses are

H_0: For British children aged 8 in 2004–2005, the mean reading score for girls was equal to the mean reading score for boys

H_1: For British children aged 8 in 2004–2005, the mean reading score for girls was not equal to the mean reading score for boys.

Solution to Exercise 4

(a) For Population A, the six different samples of size 2 with their sample means are listed below:

Sample: 10 20; sample mean $= \frac{10+20}{2} = \frac{30}{2} = 15$

Sample: 10 30; sample mean $= \frac{10+30}{2} = \frac{40}{2} = 20$

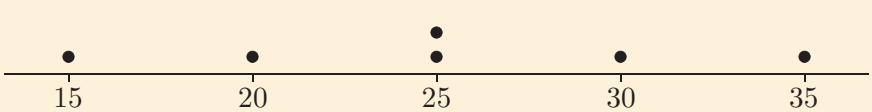
Sample: 10 40; sample mean $= \frac{10+40}{2} = \frac{50}{2} = 25$

Sample: 20 30; sample mean $= \frac{20+30}{2} = \frac{50}{2} = 25$

Sample: 20 40; sample mean $= \frac{20+40}{2} = \frac{60}{2} = 30$

Sample: 30 40; sample mean $= \frac{30+40}{2} = \frac{70}{2} = 35$

The sample means are plotted along the horizontal axis in the following figure.



Plot of values of sample means from Population A

(b) For Population B, the six different samples of size 2 with their sample means are listed below:

$$\text{Sample: 10 38; sample mean} = \frac{10+38}{2} = \frac{48}{2} = 24$$

$$\text{Sample: 10 39; sample mean} = \frac{10+39}{2} = \frac{49}{2} = 24.5$$

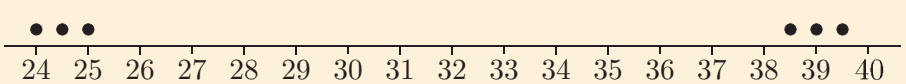
$$\text{Sample: 10 40; sample mean} = \frac{10+40}{2} = \frac{50}{2} = 25$$

$$\text{Sample: 38 39; sample mean} = \frac{38+39}{2} = \frac{77}{2} = 38.5$$

$$\text{Sample: 38 40; sample mean} = \frac{38+40}{2} = \frac{78}{2} = 39$$

$$\text{Sample: 39 40; sample mean} = \frac{39+40}{2} = \frac{79}{2} = 39.5$$

The sample means are plotted along the horizontal axis in the following figure.



Plot of values of sample means from Population B

(c) The points in the graph in part (a) are symmetrically distributed around a central mode, while the points in the graph in part (b) are split into two groups some distance apart. Hence the graph in part (a) seems more bell-shaped than the graph in part (b). This happens because the points in Population A are more symmetric – and more evenly spread out – than the points in Population B, which consist of three points close together (38, 39 and 40) and another far away (10).

Solution to Exercise 5

The distribution of reading scores – 'sample means' when $n = 1$ – is very jagged, but if you squint your eyes you get an impression of a fairly symmetric distribution with perhaps a slight preponderance of low, as opposed to high, values. The distribution of sample means of size $n = 2$ is smoother, though still with some jaggedness towards its right-hand side, fairly close to symmetric but with a little bit of left skewness. When $n = 3$ the distribution is smoother again, and any lack of symmetry is pretty small. It is also clear that the vertical scale of the sampling distribution of the mean when $n = 3$ is larger than the vertical scale of the distribution of the data ($n = 1$). By the time $n = 10$, the distribution of sample means is very smooth, symmetric, bell-shaped/normal-like and with a larger vertical scale still.

So, again, we see that even though the population distribution is not especially normal-like, as the sample size n increases, the sampling distribution of the mean quite quickly becomes much more normal-like.

Solution to Exercise 6

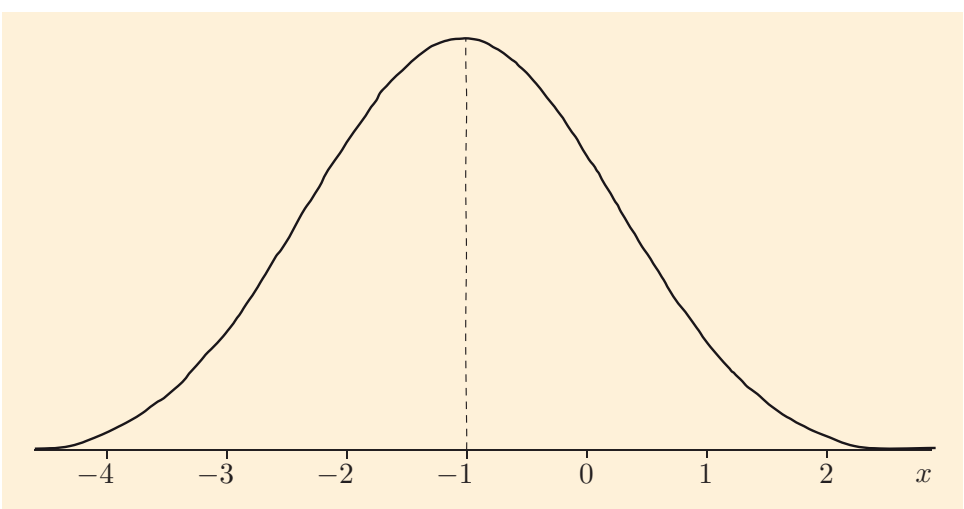
The mode of this normal distribution occurs at about $x = 10$. So $\mu \simeq 10$. Almost all the distribution is contained between $x = 0$ and $x = 20$ (i.e. within 10 ± 10). So $3\sigma \simeq 10$ and $\sigma \simeq 3.33$. That is, the normal distribution plotted in Figure 28 is approximately the normal distribution with mean $\mu = 10$ and standard deviation $\sigma = 3.33$.

Solution to Exercise 7

The mode of this normal distribution occurs at about $x = -5$. So $\mu \simeq -5$. Almost all the distribution is contained between $x = -12$ and $x = 2$ (i.e. within -5 ± 7). So $3\sigma \simeq 7$ and $\sigma \simeq 2.33$. That is, the normal distribution plotted in Figure 29 is approximately the normal distribution with mean $\mu = -5$ and standard deviation $\sigma = 2.33$.

Solution to Exercise 8

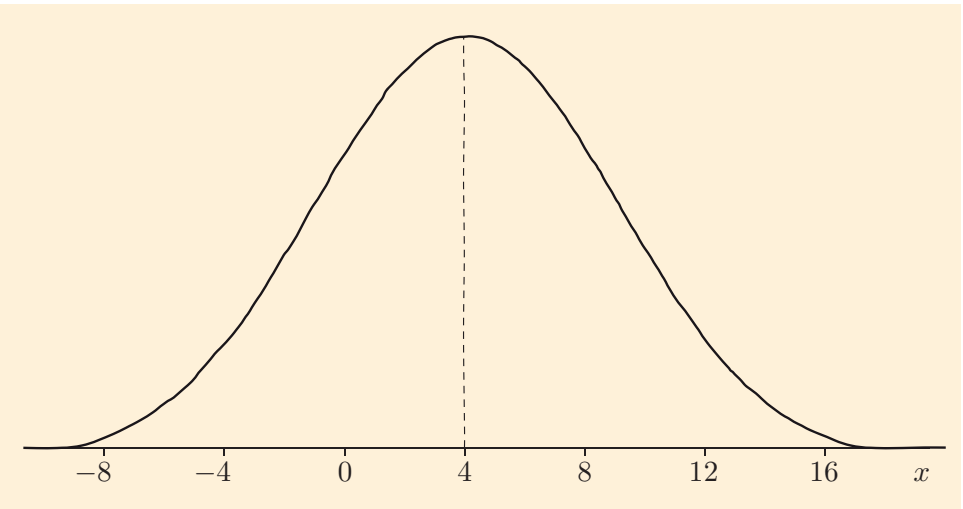
You should have obtained something like the sketch in the figure below, although, since you may have used different scales, yours could look a bit different. This normal distribution is centred at $\mu = -1$ and has just about all the distribution contained within $-1 \pm (3 \times 1) = -1 \pm 3$, i.e. between -4 and 2.



A normal distribution with $\mu = -1$, $\sigma = 1$

Solution to Exercise 9

You should have obtained something like the sketch in the figure below, although, since you may have used different scales, yours could look a bit different. This normal distribution is centred at $\mu = 4$ and has just about all the distribution contained within $4 \pm (3 \times 4) = 4 \pm 12$, i.e. between -8 and 16.



A normal distribution with $\mu = 4$, $\sigma = 4$

Solution to Exercise 10

(a) Here $\mu = 6$ and $\sigma = 3.3$, so

$$z = \frac{x - 6}{3.3}.$$

(b) Here $\mu = -6$ and $\sigma = 2$, so

$$z = \frac{x - (-6)}{2} = \frac{x + 6}{2}.$$

Solution to Exercise 11

The appropriate formula is

$$z = \frac{x - 2}{10}.$$

When $x = 3$,

$$z = \frac{3 - 2}{10} = \frac{1}{10} = 0.1.$$

Solution to Exercise 12

The appropriate formula is

$$z = \frac{x - (-1)}{0.5} = \frac{x + 1}{1/2} = 2(x + 1).$$

When $x = 0$,

$$z = 2(0 + 1) = 2.$$

Solution to Exercise 13

(a) When $n = 9$,

$$\frac{\sigma}{\sqrt{n}} = \frac{283}{\sqrt{9}} = \frac{283}{3} \simeq 94.3.$$

(b) When $n = 25$,

$$\frac{\sigma}{\sqrt{n}} = \frac{283}{\sqrt{25}} = \frac{283}{5} = 56.6.$$

(c) When $n = 100$,

$$\frac{\sigma}{\sqrt{n}} = \frac{283}{\sqrt{100}} = \frac{283}{10} = 28.3.$$

Solution to Exercise 14

(a) When $n = 4$,

$$\frac{\sigma}{\sqrt{n}} = \frac{3.6}{\sqrt{4}} = \frac{3.6}{2} = 1.8.$$

(b) When $n = 19$,

$$\frac{\sigma}{\sqrt{n}} = \frac{3.6}{\sqrt{19}} \simeq 0.83.$$

(c) When $n = 300$,

$$\frac{\sigma}{\sqrt{n}} = \frac{3.6}{\sqrt{300}} \simeq 0.21.$$

Solution to Exercise 15

When $n = 40$ and $\sigma = 0.01$,

$$\frac{\sigma}{\sqrt{n}} = \frac{0.01}{\sqrt{40}} \simeq 0.0016.$$

It follows that the sampling distribution of the mean for samples of 40 one-litre bottles of water from this manufacturer is approximately normal with mean $\mu = 1.01$ litres and standard deviation $\sigma/\sqrt{n} = 0.0016$ litres.

Solution to Exercise 16

The appropriate null and alternative hypotheses are

H_0: $\mu = 96$

H_1: $\mu \neq 96$,

where μ is the population mean reading score of all British 7-year-old girls in 2004–2005.

As the sample size, $n = 190$, is much greater than 25, it is appropriate to apply the z-test. We have

$$A = 96, \quad \overline{x} = 113.42, \quad n = 190, \quad s = 25.464.$$

The test statistic is

$$z = \frac{\overline{x} - A}{\text{ESE}} = \frac{\overline{x} - A}{s/\sqrt{n}} = \frac{113.42 - 96}{25.464/\sqrt{190}} \simeq 9.43.$$

The critical values are 1.96, -1.96 (5%) and 2.58, -2.58 (1%). Since $9.43 \geq 2.58$, we can reject the null hypothesis at the 1% significance level.

Hence there is strong evidence that the mean reading score of 7-year-old girls in 2004–2005 is not equal to 96.

This result corresponds to the similar result observed for all 7-year-old children (not just girls) in Example 7.

Solution to Exercise 17

The null and alternative hypotheses are

$$H_0\colon\ \mu = 80$$
$$H_1\colon\ \mu \neq 80,$$

where μ is the mean weight (in kg) of this breed of pig when fed the special diet.

As the sample size, 533, is greater than 25, it is appropriate to apply the z-test. We have

$$A = 80, \quad \overline{x} = 81.92, \quad n = 533, \quad s = 15.65.$$

The test statistic is

$$z = \frac{\overline{x} - A}{\text{ESE}} = \frac{\overline{x} - A}{s/\sqrt{n}} = \frac{81.92 - 80}{15.65/\sqrt{533}} \simeq 2.83.$$

The critical values are 1.96, -1.96 (5%) and 2.58, -2.58 (1%). Since $2.83 \geq 2.58$, the null hypothesis is rejected at the 1% significance level. There is strong evidence that the mean weight of this breed of pig when fed the special diet is not equal to 80 kg. There is strong evidence that the mean weight is higher for the special diet.

Solution to Exercise 18

(a) The null and alternative hypotheses are

$$H_0\colon\ \mu = 4$$
$$H_1\colon\ \mu \neq 4,$$

where μ is the population mean drying time in hours of the manufacturers' paint. As the sample size, $n = 40$, is greater than 25, it is appropriate to apply the z-test. We have

$$A = 4, \quad \overline{x} = 3.80, \quad n = 40, \quad s = 0.55.$$

The test statistic is

$$z = \frac{\overline{x} - A}{\text{ESE}} = \frac{\overline{x} - A}{s/\sqrt{n}} = \frac{3.80 - 4}{0.55/\sqrt{40}} \simeq -2.30.$$

The critical values are 1.96, -1.96 (5%) and 2.58, -2.58 (1%). Since $-2.58 < -2.30 \leq -1.96$, we can reject the null hypothesis at the 5%, though not at the 1%, significance level. We conclude that there is moderate evidence that the drying time given by the consumer magazine is incorrect. The manufacturers' paint appears to dry more quickly than the magazine claimed.

(b) You might think that such 'marginal' (moderate) evidence is not enough to conclude that the manufacturers' paint dries more quickly than the consumer magazine claimed.

The time at which paint is declared 'dry' is not well-defined: different customers might measure drying time differently or have different ideas about what 'dry' means.

Even if the measures are reliable and the test result is correct, 0.20 hours or 12 minutes is not a very large reduction. Most customers would not consider this an important difference.

You may have thought of other reservations.

Solution to Exercise 19

The null and alternative hypotheses are

$$H_0\colon\ \mu_1 = \mu_2$$
$$H_1\colon\ \mu_1 \neq \mu_2,$$

where μ_1 and μ_2 are the population mean reading scores of interest. Here and below, '1' denotes quantities relating to children with father's occupation coded 1 and '2' denotes quantities relating to children with father's occupation coded 2. The summary statistics are:

$$\overline{x}_1 = 120.55, \quad \overline{x}_2 = 117.17, \quad n_1 = 316, \quad n_2 = 203,$$
$$s_1 = 24.221, \quad s_2 = 30.085.$$

Both $n_1 = 316$ and $n_2 = 203$ are greater than 25, so we can assume that the z-test is applicable.

The estimated standard error is

$$\text{ESE} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{24.221^2}{316} + \frac{30.085^2}{203}} \simeq 2.513,$$

and the test statistic is

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\text{ESE}} \simeq \frac{120.55 - 117.17}{2.513} \simeq 1.35.$$

(You might have got 1.34, correct to two decimal places, if calculating z all in one go. Such a difference doesn't matter.)

The critical values are 1.96, -1.96 (5%) and 2.58, -2.58 (1%). Since $-1.96 < 1.35 < 1.96$, we cannot reject H_0 at the 5% significance level. There is little evidence that the mean reading score of children whose father had an occupation coded 1 differed from the mean reading score of children whose father had an occupation coded 2. This goes against conventional wisdom, at least from other contexts.

Solution to Exercise 20

Let 'A' denote quantities relating to breast-fed babies and 'B' denote quantities relating to bottle-fed babies. The null and alternative hypotheses are

$$H_0\colon\ \mu_A = \mu_B$$
$$H_1\colon\ \mu_A \neq \mu_B,$$

where μ_A and μ_B are the population mean serum calcium levels of interest. The summary statistics are:

$\overline{x}_A = 2.45, \quad \overline{x}_B = 2.30, \quad n_A = 64, \quad n_B = 169,$
$s_A = 0.292, \quad s_B = 0.274.$

Both $n_A = 64$ and $n_B = 169$ are greater than 25, so we can assume that the z-test is applicable.

The estimated standard error is

$$\text{ESE} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} = \sqrt{\frac{0.292^2}{64} + \frac{0.274^2}{169}} \simeq 0.042,$$

and the test statistic is

$$z = \frac{\overline{x}_A - \overline{x}_B}{\text{ESE}} \simeq \frac{2.45 - 2.30}{0.042} \simeq 3.57.$$

(You might have got 3.56.)

The critical values are 1.96, −1.96 (5%) and 2.58, −2.58 (1%). Since $3.57 \geq 2.58$, we can reject the null hypothesis at the 1% significance level. There is strong evidence that the mean serum calcium level of week-old babies was different depending on whether they were breast-fed or bottle-fed. The evidence is that it was higher in those who were breast-fed.

Solution to Exercise 21

Let 'R' denote quantities relating to children from rural areas and 'U' denote quantities relating to children from urban areas. The null and alternative hypotheses are

H_0: $\mu_\text{R} = \mu_\text{U}$
H_1: $\mu_\text{R} \neq \mu_\text{U}$,

where μ_R and μ_U are the population mean peak flow rates of interest (in litres per minute). The summary statistics are:

$\overline{x}_\text{U} = 226, \quad \overline{x}_\text{R} = 231, \quad n_\text{U} = 485, \quad n_\text{R} = 637,$
$s_\text{U} = 52, \quad s_\text{R} = 53.$

Both $n_\text{U} = 485$ and $n_\text{R} = 637$ are greater than 25, so we can assume that the z-test is applicable.

The estimated standard error is

$$\text{ESE} = \sqrt{\frac{s_\text{R}^2}{n_\text{R}} + \frac{s_\text{U}^2}{n_\text{U}}} = \sqrt{\frac{53^2}{637} + \frac{52^2}{485}} \simeq 3.160,$$

and the test statistic is

$$z = \frac{\overline{x}_\text{R} - \overline{x}_\text{U}}{\text{ESE}} \simeq \frac{231 - 226}{3.160} \simeq 1.58.$$

The critical values are 1.96, −1.96 (5%) and 2.58, −2.58 (1%). Since $-1.96 < 1.58 < 1.96$, we cannot reject the null hypothesis at the 5% significance level. Thus, there is little evidence to suggest that the mean peak flow rate differs between children who live in rural areas and those who live in urban areas.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Subsection 1.1 photo (a reading class): Christopher Futcher / www.istockphoto.com

Subsection 1.2 figure (*Changing Britain, Changing Lives*): www.amazon.co.uk/Changing-Britain-Lives-Generations-Century/dp/0854736506

Subsection 1.2 figure (BAS 3 documentation): www.gl-assessment.co.uk

Section 2 photo (Big Ben), taken from: www.freewebs.com/mybellringing/famousbigbells.htm

Section 3 cartoon (statistics in Greece): www.causeweb.org

Subsection 3.2 figure ('abnormally normal or normally abnormal'), taken from: www.cduniverse.com/productioninfo.asp?pid=7362503

Subsection 5.2 ESE logo: www.citrus.k12.fl.us/ese/default.htm

Section 6 ESE logo ('a more relaxing form'): Easy Serving Espresso, www.easyservingespresso.co.uk

Section 8 photo (reserved sign) © Otnaydur / Dreamstime.com

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked the publishers will be pleased to make the necessary arrangements at the first opportunity.

Index

ISBN 978-1-4730-0307-1

9 781473 003071

Cover image: minxlj/www.flickr.com